

Article

A Deep Modality-Specific Ensemble for Improving Pneumonia Detection in Chest X-rays

Sivaramakrishnan Rajaraman ^{*,†} , Peng Guo [†], Zhiyun Xue and Sameer K. Antani 

Computational Health Research Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; peng.guo@nih.gov (P.G.); zhiyun.xue@nih.gov (Z.X.); santani@mail.nih.gov (S.K.A.)

* Correspondence: sivaramakrishnan.rajaraman@nih.gov

† These authors contributed equally to this work.

Abstract: Pneumonia is an acute respiratory infectious disease caused by bacteria, fungi, or viruses. Fluid-filled lungs due to the disease result in painful breathing difficulties and reduced oxygen intake. Effective diagnosis is critical for appropriate and timely treatment and improving survival. Chest X-rays (CXRs) are routinely used to screen for the infection. Computer-aided detection methods using conventional deep learning (DL) models for identifying pneumonia-consistent manifestations in CXRs have demonstrated superiority over traditional machine learning approaches. However, their performance is still inadequate to aid in clinical decision-making. This study improves upon the state of the art as follows. Specifically, we train a DL classifier on large collections of CXR images to develop a CXR modality-specific model. Next, we use this model as the classifier backbone in the RetinaNet object detection network. We also initialize this backbone using random weights and ImageNet-pretrained weights. Finally, we construct an ensemble of the best-performing models resulting in improved detection of pneumonia-consistent findings. Experimental results demonstrate that an ensemble of the top-3 performing RetinaNet models outperformed individual models in terms of the mean average precision (mAP) metric (0.3272, 95% CI: (0.3006,0.3538)) toward this task, which is markedly higher than the state of the art (mAP: 0.2547). This performance improvement is attributed to the key modifications in initializing the weights of classifier backbones and constructing model ensembles to reduce prediction variance compared to individual constituent models.

Keywords: chest X-ray; deep learning; modality-specific knowledge; object detection; RetinaNet; ensemble learning; pneumonia; mean average precision



Citation: Rajaraman, S.; Guo, P.; Xue, Z.; Antani, S.K. A Deep Modality-Specific Ensemble for Improving Pneumonia Detection in Chest X-rays. *Diagnostics* **2022**, *12*, 1442. <https://doi.org/10.3390/diagnostics12061442>

Academic Editor: Henk A. Marquering

Received: 17 May 2022

Accepted: 8 June 2022

Published: 11 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Pneumonia is an acute respiratory infectious disease that can be caused by various pathogens such as bacteria, fungi, or viruses [1]. The infection affects the alveoli in the lungs by filling them up with fluid or pus, thereby resulting in reduced intake of oxygen and causing difficulties in breathing. The potency of the disease depends on several factors including age, health, and the source of infection. According to the World Health Organization (WHO) report (<https://www.who.int/news-room/fact-sheets/detail/pneumonia>, accessed on 11 December 2021), pneumonia is reported to be an infectious disease that results in a higher mortality rate, particularly in children. About 22% of all deaths in pediatrics from 1 to 5 years of age are reported to result from this infection. Effective diagnosis and treatment of pneumonia are therefore critical to improving patient care and survival rate.

Chest X-rays (CXRs) are commonly used to screen for pneumonia infection [2,3]. Analysis of CXR images can be particularly challenging in low and middle-income countries due to a lack of expert resources, socio-economic factors, etc. [4]. Computer-aided detection systems using conventional deep learning (DL) methods, a sub-class of machine learning (ML) algorithms can alleviate this burden and have demonstrated superiority over traditional machine learning methods in detecting disease regions of interest (ROIs) [5,6]. Such

algorithms (i) automatically detect pneumonia-consistent manifestations on CXRs; and (ii) can support clinical-decision making by facilitating swift referrals for critical cases to improve patient care.

1.1. Related Works

A study of the literature reveals several studies that propose automated methods using DL models for detecting pneumonia-consistent manifestations on CXRs. However, DL models vary in their architecture and learn discriminative features from different regions in the feature space. They are observed to be highly sensitive to data fluctuations resulting in poor generalizability due to varying degrees of biases and variances. An approach to achieving a low bias and variance and ensuring reliable outcomes is using ensemble learning which is an established ML paradigm that combines predictions from multiple diverse DL models and improves performance compared to individual constituent models [7]. The authors of [8] proposed an ensemble of FasterRCNN [9], Yolov5 [8], and EfficientDet [8] models to localize and predict bounding boxes containing pneumonia-consistent findings in the publicly available VinDr-CXR [8] dataset and reported a mean Average Precision (mAP) of 0.292. The following methods used ensembled object detection models to detect pneumonia-consistent findings using the CXR collection hosted for the RSNA Kaggle pneumonia detection challenge (<https://www.kaggle.com/c/rsna-pneumonia-detection-challenge> accessed on 3 March 2022). The current state-of-the-art method according to the challenge leaderboard (<https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/leaderboard> accessed on 3 March 2022) has a mAP of 0.2547. In [10], an ensemble of RetinaNet [11] and Mask RCNN models with ResNet-50 and ResNet-101 classifier backbones delivered a performance with a mAP of 0.2283 using the RSNA Kaggle pneumonia detection challenge CXR dataset. Another study [12] proposed a weighted-voting ensemble of the predictions from Mask R-CNN and RetinaNet models to achieve an mAP of 0.2174 in detecting pneumonia-consistent manifestations. These studies used the randomized test set split from the challenge-provided training data. This is a serious concern since the organizers have not made the blinded test set used during the challenge available for further use. This cripples follow-on research, such as ours, from making fair comparisons.

1.2. Rationale for the Study

All above studies used off-the-shelf DL object detection models with ImageNet [13] pretrained classifier backbones. However, ImageNet is a collection of stock photographic images whose visual characteristics, including shape and texture among others, are distinct from CXRs. As well, the disease-specific ROIs in CXRs are relatively small and many go unnoticed which may result in suboptimal predictions [14]. Our prior works and other literature have demonstrated that the knowledge transferred from DL models that are retrained on a large collection of CXR images is shown to improve performance on relevant target medical visual recognition tasks [15–17]. To the best of our knowledge, we observed that no literature discussed the use of CXR modality-specific backbones in object detection models, particularly applied to detecting pneumonia-consistent findings in CXRs.

1.3. Contributions of the Study

Our study improves upon the state-of-the-art as follows:

- (i). To the best of our knowledge, this is the first study that studies the impact of using CXR modality-specific classifier backbones in a RetinaNet-based object detection model, particularly applied to detecting pneumonia-consistent findings in CXRs.
- (ii). We train state-of-the-art DL classifiers on large collections of CXR images to develop CXR modality-specific models. Next, we use these models as the classifier backbone in the RetinaNet object detection network. We also initialize this backbone using random weights and ImageNet-pretrained weights to compare detection performance.

Finally, we construct an ensemble of the aforementioned models resulting in improved detection of pneumonia-consistent findings.

- (iii). Through this approach, we aim to study the combined benefits of various weight initializations for classifier backbones and construct an ensemble of the best-performing models to improve detection performance. The models' performance is evaluated in terms of mAP and statistical significance is reported in terms of confidence intervals (CIs) and p -values.

Section 2 discusses the datasets, model architecture, training strategies, loss functions, evaluation metrics, statistical methods, and computational resources, Section 3 elaborates on the results and Section 4 concludes this study.

2. Materials and Methods

2.1. Data Collection and Preprocessing

The following data collections are used for this study:

- (i). CheXpert CXR [18]: The dataset includes 223,648 frontal and lateral CXR images that are collected from 65,240 patients at Stanford Hospital, California, USA. The CXRs are labeled for 14 cardiopulmonary disease manifestations, the details are extracted from the associated radiology reports using an automated labeling algorithm.
- (ii). TBX11K CXR [19]: This collection includes 11,200 CXRs collected from normal patients and those with other cardiopulmonary abnormalities. The abnormal CXRs are collected from patients tested with the microbiological gold standard. There are 5000 CXRs showing no abnormalities and 6200 CXRs showing other abnormal findings including those collected from sick patients ($n = 5000$), active Tuberculosis (TB) ($n = 924$), latent Tuberculosis ($n = 212$), active and latent TB ($n = 54$), and other uncertain ($n = 10$) cases. The regions showing TB-consistent manifestations are labeled for the abnormal regions using coarse rectangular bounding boxes.
- (iii). RSNA CXR [20]: This CXR collection is released by RSNA for the RSNA Kaggle Pneumonia detection challenge. The collection consists of 26,684 CXRs that include 6012 CXR images showing pneumonia-consistent manifestations, 8851 CXRs showing no abnormal findings, and 11,821 CXRs showing other cardiopulmonary abnormalities. The CXRs showing pneumonia-consistent findings are labeled for abnormal regions using rectangular bounding boxes and are made available for the detection challenge.

We used the frontal CXRs from the CheXpert and TBX11K data collection during CXR image modality-specific retraining and those from the RSNA CXR collection to train the RetinaNet-based object detection models. All images are resized to 512×512 spatial dimensions to reduce computation complexity. The contrast of the CXRs is further increased by saturating the top 1% and bottom 1% of all the image pixel values. For CXR modality-specific retraining, the frontal CXR projections from the CheXpert and TBX11K datasets are divided at the patient level into 70% for training, 10% for validation, and 20% for testing. This patient-level split prevents the leakage of data and subsequent bias during model training. For object detection, the frontal CXRs from the RSNA CXR dataset that shows pneumonia-consistent manifestations are divided at the patient level into 70% for training, 10% for validation, and 20% for testing. Table 1 shows the number of CXR images across the training, validation, and test sets used for CXR modality-specific retraining and object detection, respectively.

Table 1. Patient-level dataset splits show the number of images for CXR modality-specific retraining and object detection. Note: TBX11K and RSNA datasets have one image per patient.

Dataset	Train		Validation		Test	
	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
CXR Modality-specific retraining						
CheXpert	13,600	13,600	1700	1700	1700	1700
TBX11k	3040	3040	380	380	380	380
RetinaNet-based object detection						
Dataset	Train		Validation		Test	
RSNA	4212		600		1200	

2.2. Model Architecture

2.2.1. CXR Modality-Specific Retraining

The ImageNet-pretrained DL models, viz., VGG-16, VGG-19, DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNet have demonstrated promising performance in several medical visual recognition tasks [14,19,21–23]. These models are further retrained on a large collection of CXR images to classify them as showing cardiopulmonary abnormal manifestations or no abnormalities. Such retraining helps the models to learn CXR image modality-specific features that can be transferred and fine-tuned to improve performance in a relevant task using CXR images. The best-performing model with the learned CXR image modality-specific weights is used as the classifier backbone to train the RetinaNet-based object detection model toward detecting pneumonia-consistent manifestations. Figure 1 shows the block diagram illustrating the steps involved in CXR image modality-specific retraining.

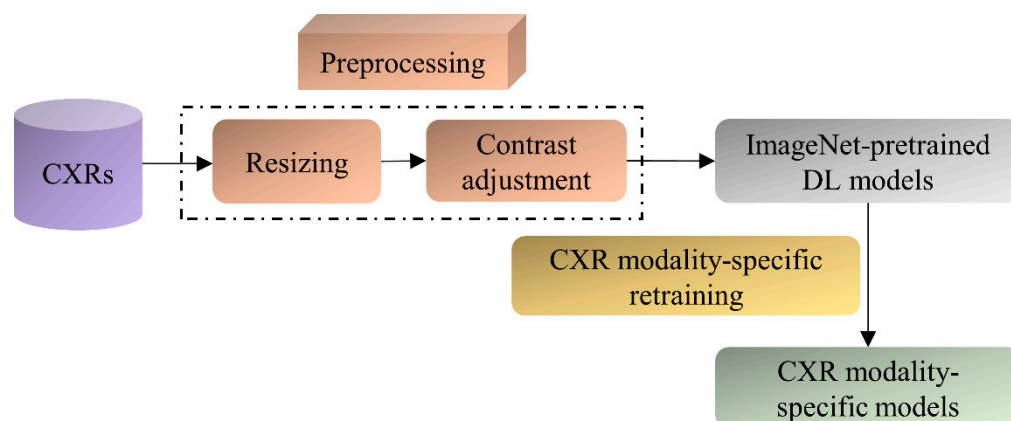


Figure 1. Steps illustrating CXR image modality-specific retraining of the ImageNet-pretrained models.

2.2.2. RetinaNet Architecture

We used RetinaNet as the base object detection architecture in our experiments. The architecture of the RetinaNet model is shown in Figure 2. As a single-stage object detection structure, RetinaNet shares a similar concept of “anchor proposal” with [24]. It used a feature pyramid network (FPN) [25] where features on each of the image scales are computed separately in the lateral connections and then summed up through convolutional operations via the top-down pathways. The FPN network combines low-resolution features with strong semantic information, and high-resolution features with weak semantics through top-down paths and horizontal connections. Thus, feature maps with rich semantic information are obtained that would prove beneficial for detecting relatively smaller ROIs consistent with pneumonia compared to the other parts of the CXR image. Furthermore, when trained to minimize the focal loss [5], the RetinaNet was reported to deliver significant performance focusing on hard, misclassified examples.

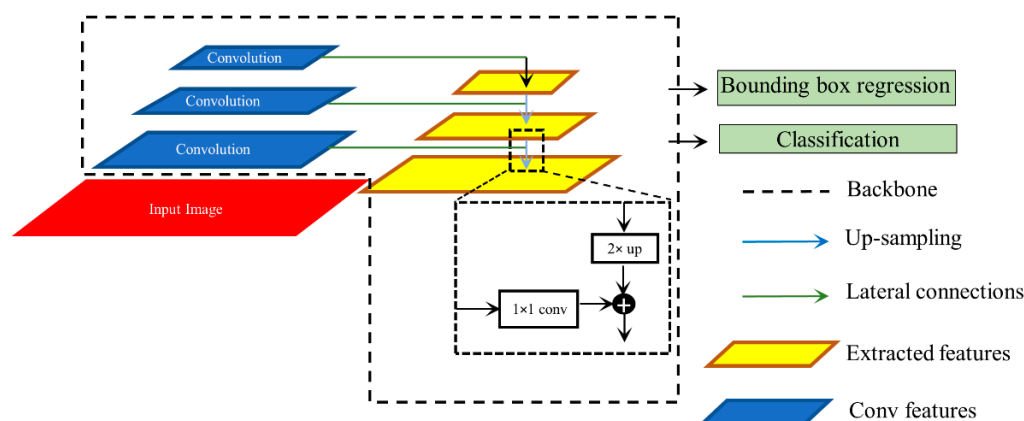


Figure 2. Method flowchart for the RetinaNet network.

2.2.3. Ensemble of RetinaNet Models with Various Backbones

We initialized the weights of the VGG-16 and ResNet-50 classifier backbones used in the RetinaNet model using three strategies: (i) Random weights; (ii) ImageNet-pretrained weights, and (iii) CXR image modality-specific retrained weights as discussed in Section 2.2.1. Each model is trained for 80 epochs and the model weights (snapshots) are stored at the end of each epoch. Varying modifications of the RetinaNet model classifier backbones and loss functions are mentioned in Table 2.

Table 2. RetinaNet model classifier backbones with varying weight initializations and loss functions. The loss functions mentioned are used for classification. For bounding box regression, only the smooth-L1 loss function [26] is used in all cases.

ResNet-50 Backbone and Classification Loss Functions	VGG-16 Backbone and Classification Loss Functions
ResNet-50 with random weights + focal loss	VGG-16 with random weights + focal loss
ResNet-50 with random weights + focal Tversky loss	VGG-16 with random weights + focal Tversky loss
ResNet-50 with ImageNet pretrained weights + focal loss	VGG-16 with ImageNet pretrained weights + focal loss
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	VGG-16 with ImageNet pretrained weights + focal Tversky loss
ResNet-50 with CXR image modality-specific weights + focal loss	VGG-16 with CXR image modality-specific weights + focal loss
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	VGG-16 with CXR image modality-specific weights + focal Tversky loss

We adopted the non-maximum suppression (NMS) in the RetinaNet training with an IoU threshold of 0.5 and evaluated the models using all the predictions with a confidence score over 0.9. A weighted averaging ensemble is constructed using (i) the top-3 performing models from the 12 RetinaNet models mentioned in Table 2, and (ii) the top-3 performing snapshots (model weights) using each classifier backbone. We empirically assigned the weights as 1, 0.9, and 0.8 for the predictions of the 1st, 2nd, and 3rd best performing models. A schematic of the ensemble procedure is shown in Figure 3. An ensemble bounding box is generated if the IOU of the weighted average of the predicted bounding boxes and the ground truth (GT) boxes is greater than 0.5. The ensemble model is evaluated based on the mean average precision (mAP) metric.

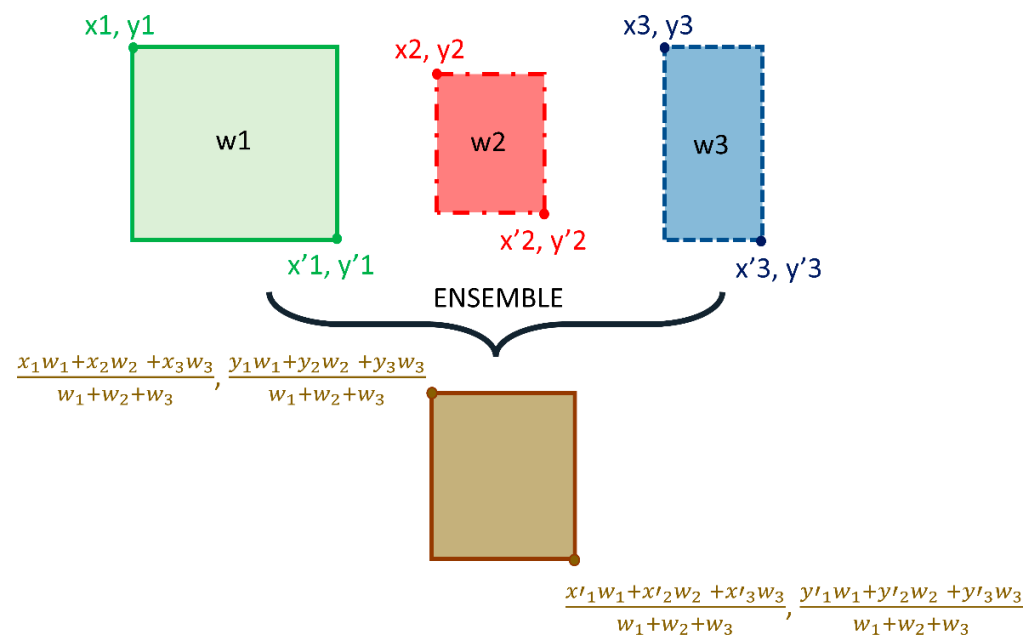


Figure 3. Method Schematic of the ensemble approach.

2.2.4. Loss Functions and Evaluation Metrics

CXR Image Modality-Specific Retraining

During CXR image modality-specific retraining, the DL models are retrained on a combined selection of the frontal CXR projections from the CheXpert and TBX11K datasets (details in Table 1). The training is performed for 128 epochs to minimize the categorical cross-entropy (CCE) loss. The CCE loss is the most commonly used loss function in classification tasks, and it helps to measure the distinguishability between two discrete probability distributions. It is expressed as shown in Equation (1).

$$CCE_{loss} = - \sum_{k=1}^{output\ size} y_k \log y_k \quad (1)$$

Here, y_k denotes the k th scalar value in the model output, y_k denotes the corresponding target, and the *output size* denotes the number of scalar values in the model output. The term y_k denotes the probability that event k occurs and the sum of all $y_k = 1$. The minus sign in the CCE loss equation ensures the loss is minimized when the distributions become less distinguishable. We used a stochastic gradient descent optimizer with an initial learning rate of 1×10^{-4} and momentum of 0.9 to reduce the CCE loss and improve performance. Callbacks are used to store the model checkpoints and the learning rate is reduced after a patience parameter of 10 epochs when the validation performance ceased to improve. The weights of the model that delivered a superior performance with the validation set are used to predict the test set. The models are evaluated in terms of accuracy, the area under the receiver-operating characteristic curve (AUROC), the area under the precision-recall (PR) curve (AUPRC), sensitivity, precision, F-score, Matthews correlation coefficient (MCC), and Kappa statistic.

RetinaNet-Based Detection of Pneumonia-Consistent Findings

Considering medical images, the disease ROIs span a relatively smaller portion of the whole image. This results in a considerably high degree of imbalance in the foreground ROI and the background pixels. These issues are particularly prominent in applications such as detecting cardiopulmonary manifestations like pneumonia where the number of pixels showing pneumonia-consistent manifestations is markedly lower compared to the total number of image pixels. Generalized loss functions such as balanced cross-entropy loss do

not take this data imbalance into account. This may lead to a learning bias and subsequent adversity in learning the minority ROI pixels. Appropriate selection of the loss function is therefore critical for improving detection performance. In this regard, the authors of [11] proposed the focal loss for object detection, an extension of the cross-entropy loss, which alleviates this learning bias by giving importance to the minority ROI pixels while down-weighting the majority background pixels. Minimizing the focal loss thereby reduces the loss contribution from majority background examples and increases the importance of correctly detecting the minority disease-positive ROI pixels. The focal loss is expressed as shown in Equation (2).

$$Focal\ loss(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{2}$$

Here, p_t denotes the probability the object detection model predicts for the GT. The parameter γ decides the rate of down-weighting the majority (background non-ROI) samples. The equation converges to the conventional cross-entropy loss when $\gamma = 0$. We empirically selected the value of $\gamma = 2$ which delivered superior detection performance.

Another loss function called the Focal Tversky loss function [27], a generalization of the focal loss function, is proposed to tackle the data imbalance problem and is given in Equation (3). The Focal Tversky loss function generalizes the Tversky loss which is based on the Tversky index that helps achieve a superior tradeoff between recall and precision when trained on class-imbalanced datasets. The Focal Tversky loss function uses a smoothing parameter γ that controls the non-linearity of the loss at different values of the Tversky index to balance between the minority pneumonia-consistent ROI and majority background classes. In Equation (3), TI denotes the Tversky index, expressed as shown in Equation (4).

$$FT_{loss_c} = \sum_c 1 - TI_c^\gamma \tag{3}$$

$$TI_c = \frac{\sum_{i=1}^M t_{ic}g_{ic} + \epsilon}{\sum_{i=1}^M t_{ic}g_{ic} + \alpha \sum_{i=1}^M t_{i\hat{c}}g_{ic} + \beta \sum_{i=1}^M t_{ic}g_{i\hat{c}} + \epsilon} \tag{4}$$

Here, g_{ic} and t_{ic} denote the ground truth and predicted labels for the pneumonia class c , where g_{ic} and $t_{ic} \in \{0,1\}$. That is, t_{ic} denotes the probability that the pixel i belongs to the pneumonia class c and $t_{i\hat{c}}$ denotes the probability that the pixel i belongs to the background class \hat{c} . The same holds for g_{ic} and $g_{i\hat{c}}$. The term M denotes the total number of image pixels. The term ϵ provides numerical stability to avoid divide-by-zero errors. The hyperparameters α and β are tuned to emphasize recall under class-imbalanced training conditions. The Tversky index is adapted to a loss function by minimizing $\sum_c 1 - TI_c$. After empirical evaluations, we fixed the value of $\gamma = 4/3$, $\alpha = 0.7$ and $\beta = 0.75$.

As is known, the loss function within RetinaNet is a summation of a couple of loss functions, one for classification and the other for bounding box regression. We left the Smooth-L1 loss that is used for bounding box regression unchanged. For classification, we explored the performance with focal loss and focal Tversky loss functions individually for training the RetinaNet models with varying weight initializations. We used the bounding box annotations [20] associated with the RSNA CXRs showing pneumonia-consistent manifestations as the GT bounding boxes and measured its agreement with that generated by the models initialized with random weights, ImageNet-pretrained, and CXR image modality-specific retrained classifier backbones. Let TP, FP, and FN denote the true positives, false positives, and false negatives, respectively. Given a pre-defined IOU threshold, a predicted bounding box is considered to be TP if it overlaps with the GT bounding box by a value equal to or exceeding this threshold. FP denotes that the predicted bounding box has no associated GT bounding box. FN denotes the GT bounding box has no associated predicted bounding box. The mAP is measured as the area under the precision-recall curve (AUPRC) as shown in Equation (5). Here, P denotes precision which measures the accuracy of predictions, and R denotes recall which measures how well the model identifies all the

Tps. They are computed as shown in Equations (6) and (7). The value of mAP lies in the range [0, 1].

$$\text{mean average precision (mAP)} = \int_0^1 P(R) dR \quad (5)$$

$$\text{Precision (P)} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (7)$$

We used a Linux system with 1080Ti GPU, the Tensorflow backend (v. 2.6.2) with Keras, and CUDA/CUDNN libraries for accelerating the graphical processing unit (GPU) toward training the object detection models that are configured in the Python environment.

2.3. Statistical Analysis

We evaluated statistical significance using the mAP metric achieved by the models trained with various weight initializations and loss functions. The 95% confidence intervals (CIs) are measured as the binomial interval using the Clopper-Pearson method.

3. Results and Discussion

We organized the results from our experiments into the following sections: Evaluating the performance of (i) CXR image modality-specific retrained models and (ii) RetinaNet object detection models using classifier backbones with varying weight initializations and loss functions.

3.1. Classification Performance during CXR Image Modality-Specific Retraining

Recall that the ImageNet-pretrained DL models are retrained on the combined selection of CXRs from the CheXpert and TBX11K collection. Such retraining is performed to convert the weight layers specific to the CXR image modality and let the models learn CXR modality-specific features to improve performance when the learned knowledge is transferred and fine-tuned for a related medical image visual recognition task. The performance achieved by the CXR image modality-specific retrained models using the hold-out test set is listed in Table 3 and the performance curves are shown in Figure 4. The *no-skill* line in Figure 4 denotes the performance when a classifier would fail to discriminate between the normal and abnormal CXRs and therefore would predict a random outcome or a specific category under all circumstances.

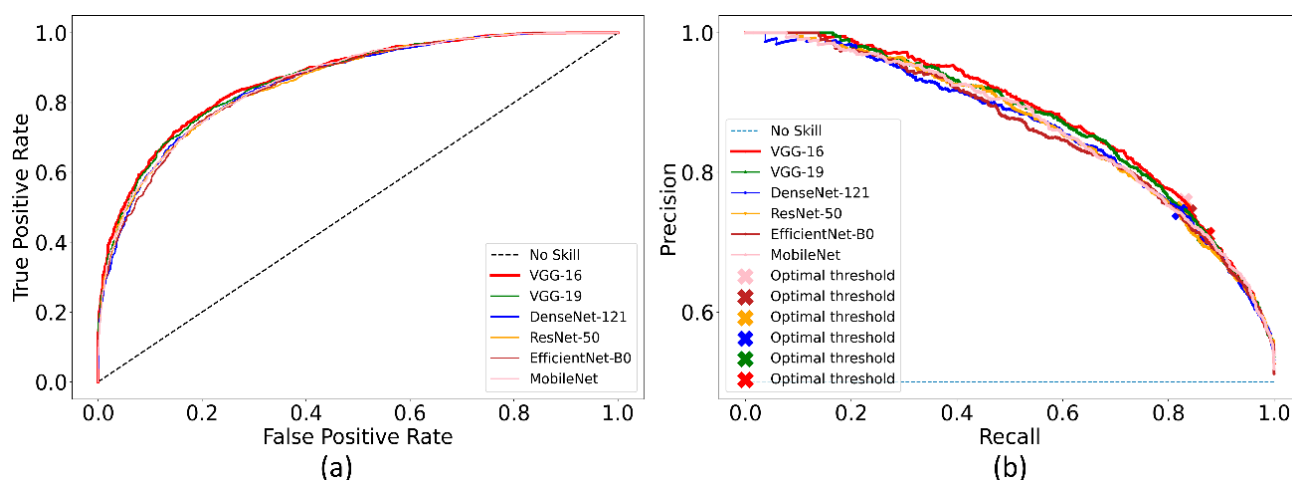


Figure 4. The collection of performance curves for the CXR image modality-specific retrained models. The performance is recorded at the optimal classification threshold measured with the validation data. (a) ROC and (b) PR curves.

Table 3. Performance of the CXR image modality-specific retrained models with the hold-out test set. Bold numerical values denote superior performance. The values in parenthesis denote the 95% CI for the MCC metric.

Models	Accuracy	AUROC	AUPRC	Sensitivity	Precision	F-Score	MCC	Kappa
VGG-16	0.7834	0.8701	0.8777	0.8303	0.7591	0.7931	0.5693 (0.5542, 0.5844)	0.5668
VGG-19	0.7743	0.8660	0.8727	0.8389	0.7429	0.7880	0.5532 (0.5380, 0.5684)	0.5486
DenseNet-121	0.7738	0.8582	0.8618	0.8264	0.7477	0.7851	0.5507 (0.5355, 0.5659)	0.5476
ResNet-50	0.7685	0.8586	0.8646	0.8207	0.7431	0.7800	0.5400 (0.5248, 0.5552)	0.5370
EfficientNet-B0	0.7553	0.8568	0.8612	0.8678	0.7084	0.7800	0.5240 (0.5088, 0.5392)	0.5106
MobileNet	0.7584	0.8609	0.8655	0.8726	0.7104	0.7832	0.5309 (0.5157, 0.5461)	0.5168

We could observe from Table 3 that the CXR image modality-specific retrained VGG-16 model demonstrates the best performance compared to other models in terms of all metrics except sensitivity. Of these, the MCC metric is a good measure to use because unlike F-score because it considers a balanced ratio of TPs TNs, FPs, and FNs. We noticed that the differences in the MCC values achieved by the various CXR image modality-specific retrained models are not significantly different ($p > 0.05$). Based on its performance, we used VGG-16 as the backbone for the RetinaNet detector. However, to enable fair comparison with other conventional RetinaNet-based results, we included the ResNet-50 backbone for detecting pneumonia-consistent manifestations. The VGG-16 and ResNet-50 classifier backbones are also initialized with random and ImageNet-pretrained weights for further comparison.

3.2. Detection Performance Using RetinaNet Models and Their Ensembles

Recall that the RetinaNet models are trained with different initializations of the classifier backbones. The performance achieved by these models using the hold-out test set is listed in Table 4. Figure 5 shows the PR curves obtained with the RetinaNet model using varying weight initializations for the selected classifier backbones. These curves show the precision and recall value of the model's bounding box predictions on every sample in the test set. We observe from Table 4 that the RetinaNet model with the CXR image modality-specific retrained ResNet-50 classifier backbone and trained using the focal loss function demonstrates superior performance in terms of mAP. Figure 6 shows the bounding box predictions of the top-3 performing RetinaNet models for a sample CXR from the hold-out test set.

We used two approaches to combine the bounding box predictions. They are (i) using the bounding box predictions from the top-3 performing RetinaNet models, viz., ResNet-50 with CXR image modality-specific weights + focal loss, ResNet-50 with CXR image modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss; and, (ii) using the bounding box predictions from the top-3 performing snapshots (weights) within each model. The results are presented in Table 5 and Figure 7. A weighted averaging ensemble of the bounding boxes is generated when the IoU of the predicted bounding boxes is greater than the threshold value which is set at 0.5. Recall that the models are trained for 80 epochs and a snapshot (i.e., the model weights) is stored at the end of each epoch. We observed that the ensemble of the top-3 performing RetinaNet models delivered superior performance in terms of mAP metric compared to other models and ensembles. Figure 8 shows a sample CXR image with GT and predicted bounding

boxes using the weighted averaging ensemble of the top-3 individual models and the top-3 snapshots of the best-performing model.

Table 4. Performance of RetinaNet with the varying weight initializations for the classifier backbones and training losses. The values in parenthesis denote the 95% CI for the mAP metric. Bold numerical values denote superior performance.

Models	AUPRC (mAP)
ResNet-50 with random weights + focal loss	0.2763 (0.2509, 0.3017)
ResNet-50 with random weights + focal Tversky loss	0.2627 (0.2377, 0.2877)
ResNet-50 with ImageNet pretrained weights + focal loss	0.2719 (0.2467, 0.2971)
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	0.2737 (0.2484, 0.2990)
ResNet-50 with CXR image modality-specific weights + focal loss	0.2865 (0.2609, 0.3121)
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	0.2859 (0.2603, 0.3115)
VGG-16 with random weights + focal loss	0.2549 (0.2302, 0.2796)
VGG-16 with random weights + focal Tversky loss	0.2496 (0.2251, 0.2741)
VGG-16 with ImageNet pretrained weights + focal loss	0.2734 (0.2481, 0.2987)
VGG-16 with ImageNet pretrained weights + focal Tversky loss	0.2666 (0.2415, 0.2917)
VGG-16 with CXR image modality-specific weights + focal loss	0.2686 (0.2435, 0.2937)
VGG-16 with CXR image modality-specific weights + focal Tversky loss	0.2648 (0.2398, 0.2898)

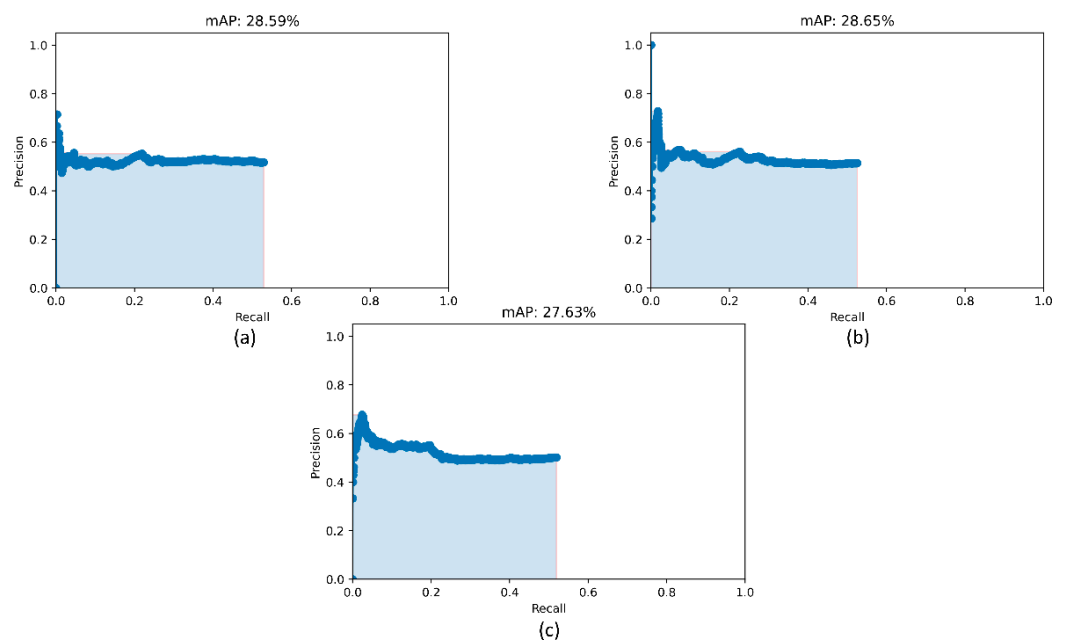


Figure 5. PR curves of the RetinaNet models initialized with varying weights for the classifier backbones. (a) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (b) ResNet-50 with CXR image modality-specific weights + focal loss, and (c) ResNet-50 with random weights + focal loss.

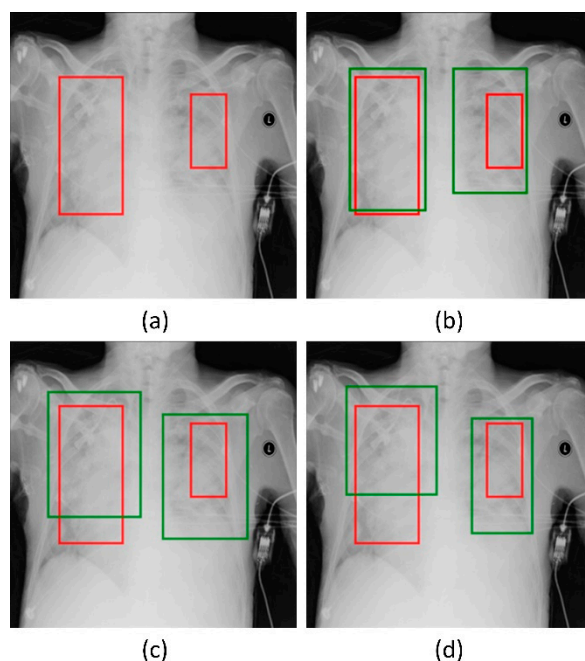


Figure 6. Bounding box predictions of the RetinaNet models initialized with varying weights for the classifier backbones. Green boxes denote the model predictions and red boxes denote the ground truth. (a) A sample CXR with ground truth bounding boxes. (b) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (c) ResNet-50 with CXR image modality-specific weights + focal loss, and (d) ResNet-50 with random weights + focal loss.

Table 5. Ensemble performance with the top-3 performing models (from Table 4) and the top-3 snapshots for each of the models trained with various classifier backbones and weight initializations. Values in parenthesis denote the 95% CI for the mAP metric. Bold numerical values denote superior performance.

Ensemble Method	mAP
Top-3 model ensemble (ResNet-50 with CXR image modality-specific weights + focal loss, ResNet-50 with CXR image modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss)	0.3272 (0.3006, 0.3538)
Ensemble of the top-3 snapshots for each model	
ResNet-50 with random weights + focal loss	0.2777 (0.2523, 0.3031)
ResNet-50 with random weights + focal Tversky loss	0.2630 (0.2380, 0.2880)
ResNet-50 with ImageNet pretrained weights + focal loss	0.2788 (0.2534, 0.3042)
ResNet-50 with ImageNet pretrained weights + focal Tversky loss	0.2812 (0.2557, 0.3067)
ResNet-50 with CXR image modality-specific weights + focal loss	0.2973 (0.2714, 0.3232)
ResNet-50 with CXR image modality-specific weights + focal Tversky loss	0.2901 (0.2644, 0.3158)
VGG-16 with random weights + focal loss	0.2633 (0.2383, 0.2883)
VGG-16 with random weights + focal Tversky loss	0.2556 (0.2309, 0.2803)
VGG-16 with ImageNet pretrained weights + focal loss	0.2823 (0.2568, 0.3078)
VGG-16 with ImageNet pretrained weights + focal Tversky loss	0.2715 (0.2463, 0.2967)
VGG-16 with CXR image modality-specific weights + focal loss	0.2813 (0.2558, 0.3068)
VGG-16 with CXR image modality-specific weights + focal Tversky loss	0.2698 (0.2446, 0.2950)

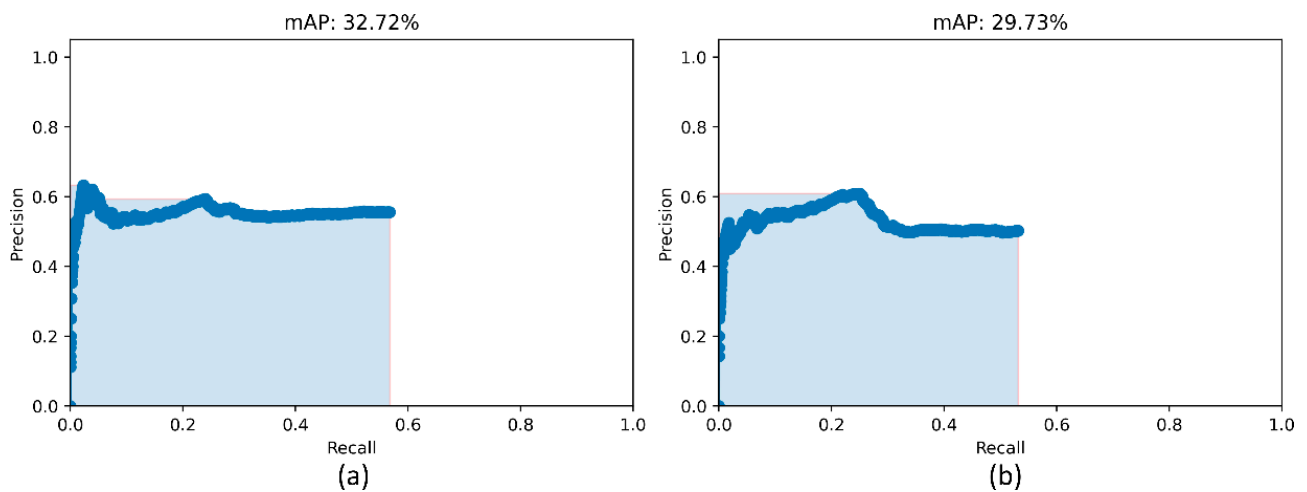


Figure 7. PR curves of the model ensembles. (a) PR curve obtained with the weighted-averaging ensemble of top-3 performing models (ResNet-50 with CXR modality-specific weights + focal loss, ResNet-50 with CXR modality-specific weights + focal Tversky loss, and ResNet-50 with random weights + focal loss) and (b) PR curve obtained with the ensemble of top-3 performing snapshots while training the ResNet-50 with CXR modality-specific weights + focal loss model.

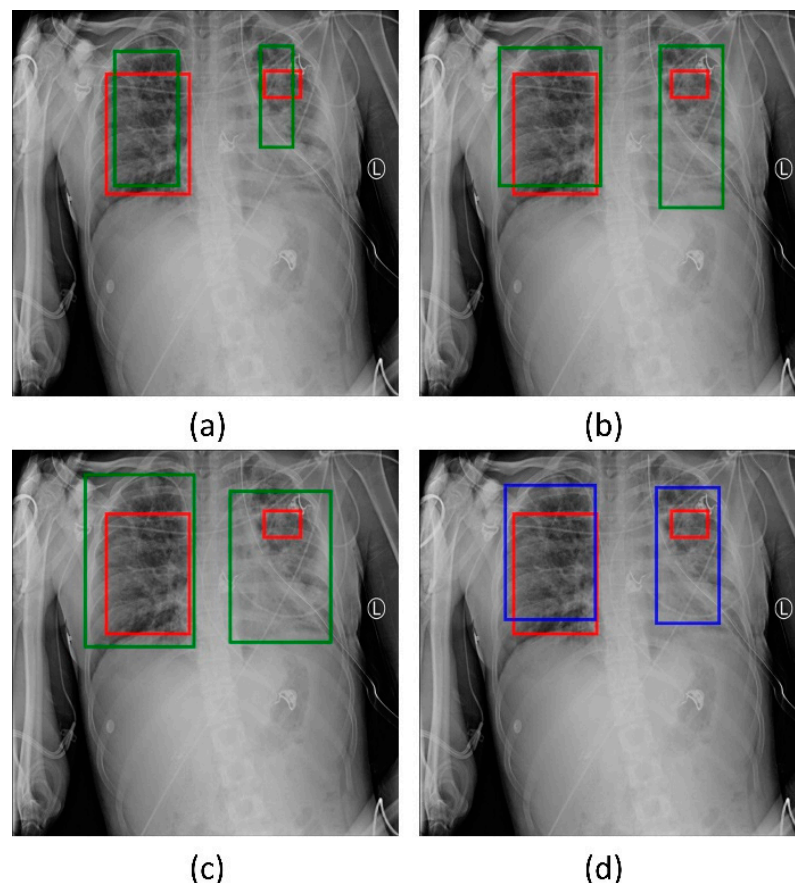


Figure 8. Bounding box predictions using the ensemble of RetinaNet models initialized with varying weights for the classifier backbones. Green boxes denote the individual model predictions, blue boxes denote the ensemble predictions and red boxes denote the ground truth. (a) ResNet-50 with CXR image modality-specific weights + focal Tversky loss; (b) ResNet-50 with CXR image modality-specific weights + focal loss; (c) ResNet-50 with random weights + focal loss, and (d) the ensemble bounding box prediction.

4. Conclusions and Future Work

In this study, we demonstrated the combined benefits of training CXR image modality-specific models, using them as backbones in an object detection model, evaluating them in different loss settings, and constructing ensembles of the best-performing models to improve performance in a pneumonia detection task. We observed that both CXR image modality-specific classifier backbones and ensemble learning improved detection performance compared to the individual constituent models. This study, however, suffers from the limitation that we have only investigated the effect of using CXR modality-specific classifier backbones in a RetinaNet-based object detection model to improve detecting pneumonia-consistent findings. The efficacy of this approach in detecting other cardiopulmonary disease manifestations is a potential avenue for future research. Additional diversity in the training process could be introduced by using CXR images and their disease-specific annotations collected from multiple institutions. With the advent of high-performance computing and current advancements in DL-based object detection, future studies could explore the use of mask x-RCNN, transformer-based models, and other advanced detection methods [28–31] and their ensembles in improving detection performance. Novel model optimization methods and loss functions can be proposed to further improve detection performance. However, the objective of this study is not to propose a new objection detection model but to validate the use of CXR modality-specific classifier backbones in existing models to improve performance. As the organizers of the RSNA Kaggle pneumonia detection challenge have not made the blinded GT annotations of the test set publicly available, we are unable to compare our results with the challenge leaderboard. However, the performance of our method on a random split from the challenge-provided training set, where we sequester 10% of the images for testing, using 70% for training and 20% for validation, respectively, is markedly superior to the best performing method on the leaderboard.

Author Contributions: Conceptualization, S.R., P.G. and S.K.A.; Data curation, S.R. and P.G.; Formal analysis, S.R., P.G. and Z.X.; Funding acquisition, S.K.A.; Investigation, Z.X. and S.K.A.; Methodology, S.R. and P.G.; Project administration, S.K.A.; Resources, S.K.A.; Software, S.R. and P.G.; Supervision, Z.X. and S.K.A.; Validation, S.R., P.G. and S.K.A.; Visualization, S.R. and P.G.; Writing—original draft, S.R. and P.G.; Writing—review & editing, S.R., P.G., Z.X. and S.K.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The funders had no role in the study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: Ethical review and approval were waived for this study because of the retrospective nature of the study and the use of anonymized patient data.

Informed Consent Statement: Patient consent was waived by the IRBs because of the retrospective nature of this investigation and the use of anonymized patient data.

Data Availability Statement: The data required to reproduce this study is publicly available and cited in the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e9. [[CrossRef](#)] [[PubMed](#)]
2. Rajaraman, S.; Candemir, S.; Thoma, G.; Antani, S. Visualizing and Explaining Deep Learning Predictions for Pneumonia Detection in Pediatric Chest Radiographs. *Proc. SPIE* **2019**, *10950*, 1095005.
3. Nishio, M.; Noguchi, S.; Matsuo, H.; Murakami, T. Automatic Classification between COVID-19 Pneumonia, Non-COVID-19 Pneumonia, and the Healthy on Chest X-Ray Image: Combination of Data Augmentation Methods. *Sci. Rep.* **2020**, *10*, 17532. [[CrossRef](#)]

4. Balabanova, Y.; Coker, R.; Fedorin, I.; Zakharova, S.; Plavinskij, S.; Krukov, N.; Atun, R.; Drobniowski, F. Variability in Interpretation of Chest Radiographs among Russian Clinicians and Implications for Screening Programmes: Observational Study. *BMJ* **2005**, *331*, 379–382. [[CrossRef](#)] [[PubMed](#)]
5. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S.; Rajaraman, S.; Candemir, S.; Kim, I.; Thoma, G.; Antani, S. Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. *Appl. Sci.* **2018**, *8*, 1715. [[CrossRef](#)]
7. Mouhafid, M.; Salah, M.; Yue, C.; Xia, K. Deep Ensemble Learning-Based Models for Diagnosis of COVID-19 from Chest CT Images. *Healthcare* **2022**, *10*, 166. [[CrossRef](#)] [[PubMed](#)]
8. Pham, V.T.; Tran, C.M.; Zheng, S.; Vu, T.M.; Nath, S. Chest X-Ray Abnormalities Localization via Ensemble of Deep Convolutional Neural Networks. *Int. Conf. Adv. Technol. Commun.* **2021**, *2021*, 125–130. [[CrossRef](#)]
9. Xie, X.; Liao, Q.; Ma, L.; Jin, X. Gated Feature Pyramid Network for Object Detection. *Lect. Notes Comput. Sci.* **2018**, *11259*, 199–208. [[CrossRef](#)]
10. Mao, L.; Yumeng, T.; Lina, C. Pneumonia detection in chest X-rays: A deep learning approach based on ensemble RetinaNet and mask R-CNN. In Proceedings of the 8th International Conference on Advanced Cloud and Big Data, Taiyuan, China, 19–20 September 2020; pp. 213–218. [[CrossRef](#)]
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [[CrossRef](#)] [[PubMed](#)]
12. Ko, H.; Ha, H.; Cho, H.; Seo, K.; Lee, J. Pneumonia detection with weighted voting ensemble of CNN models. In Proceedings of the 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 26–29 May 2019; pp. 306–310. [[CrossRef](#)]
13. Fei-Fei, L.; Deng, J.; Li, K. ImageNet: Constructing a Large-Scale Image Database. *J. Vis.* **2010**, *9*, 1037. [[CrossRef](#)]
14. Suzuki, K. Overview of Deep Learning in Medical Imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [[CrossRef](#)] [[PubMed](#)]
15. Rajaraman, S.; Sornapudi, S.; Kohli, M.; Antani, S. Assessment of an ensemble of machine learning models toward abnormality detection in chest radiographs. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019.
16. Rajaraman, S.; Folio, L.R.; Dimperio, J.; Alderson, P.O.; Antani, S.K. Improved Semantic Segmentation of Tuberculosis—Consistent Findings in Chest x-Rays Using Augmented Training of Modality-Specific u-Net Models with Weak Localizations. *Diagnostics* **2021**, *11*, 616. [[CrossRef](#)] [[PubMed](#)]
17. Yadav, O.; Passi, K.; Jain, C.K. Using deep learning to classify x-ray images of potential tuberculosis patients. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018.
18. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghighi, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc. Conf. AAAI Artif. Intell.* **2019**, *33*, 590–597. [[CrossRef](#)]
19. Liu, Y.; Wu, Y.H.; Ban, Y.; Wang, H.; Cheng, M.M. Rethinking computer-aided tuberculosis diagnosis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
20. Shih, G.; Wu, C.C.; Halabi, S.S.; Kohli, M.D.; Prevedello, L.M.; Cook, T.S.; Sharma, A.; Amorosa, J.K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol. Artif. Intell.* **2019**, *1*, e180041. [[CrossRef](#)] [[PubMed](#)]
21. Rajaraman, S.; Jaeger, S.; Thoma, G.R.; Antani, S.K.; Silamut, K.; Maude, R.J.; Hossain, M.A. Understanding the Learned Behavior of Customized Convolutional Neural Networks toward Malaria Parasite Detection in Thin Blood Smear Images. *J. Med. Imaging* **2018**, *5*, 034501. [[CrossRef](#)] [[PubMed](#)]
22. Ganesan, P.; Rajaraman, S.; Long, R.; Ghoraani, B.; Antani, S. Assessment of data augmentation strategies toward performance improvement of abnormality classification in chest radiographs. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019.
23. Rajaraman, S.; Antani, S.K.; Poostchi, M.; Silamut, K.; Hossain, M.A.; Maude, R.J.; Jaeger, S.; Thoma, G.R.; Hossain, A.; Maude, R.J.; et al. Pre-Trained Convolutional Neural Networks as Feature Extractors toward Improved Malaria Parasite Detection in Thin Blood Smear Images. *PeerJ* **2018**, *6*, e4568. [[CrossRef](#)] [[PubMed](#)]
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
25. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [[CrossRef](#)]
26. Liu, C.; Yu, S.; Yu, M.; Wei, B.; Li, B.; Li, G.; Huang, W. Adaptive smooth L1 loss: A better way to regress scene texts with extreme aspect ratios. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Athens, Greece, 5–8 September 2021. [[CrossRef](#)]
27. Abraham, N.; Khan, N.M. A Novel Focal Tversky Loss Function with Improved Attention U-Net for Lesion Segmentation. In Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019.

28. Qi, W.; Su, H. A Cybertwin Based Multimodal Network for ECG Patterns Monitoring Using Deep Learning. *IEEE Trans. Ind. Inform.* **2022**, *3203*, 1–9. [[CrossRef](#)]
29. Su, H.; Hu, Y.; Karimi, H.R.; Knoll, A.; Ferrigno, G.; De Momi, E. Improved Recurrent Neural Network-Based Manipulator Control with Remote Center of Motion Constraints: Experimental Results. *Neural Netw.* **2020**, *131*, 291–299. [[CrossRef](#)]
30. Qi, W.; Aliverti, A. A Multimodal Wearable System for Continuous and Real-Time Breathing Pattern Monitoring during Daily Activity. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 2199–2207. [[CrossRef](#)]
31. Su, H.; Mariani, A.; Ovrur, S.E.; Menciassi, A.; Ferrigno, G.; De Momi, E. Toward Teaching by Demonstration for Robot-Assisted Minimally Invasive Surgery. *IEEE Trans. Autom. Sci. Eng.* **2021**, *18*, 484–494. [[CrossRef](#)]