Interpreting Deep Ensemble Learning through Radiologist Annotations for COVID-19 Detection in Chest Radiographs

Sivaramakrishnan Rajaraman[1*], Sudhir Sornapudi[2], Philip O Alderson[3], Les R Folio[4],
Sameer K Antani[1]

1 Lister Hill National Center for Biomedical Communications, National Library of
Medicine, Maryland, Bethesda, United States of America
2 Department of Electrical and Computer Engineering, Missouri University of Science
and Technology, Missouri, Rolla, United States of America
3 School of Medicine, Saint Louis University, Missouri, St. Louis, United States of
America
4 Radiological and Imaging Sciences, Clinical Center, National Institutes of Health,
Maryland, Bethesda

* email:sivaramakrishnan.rajaraman@nih.gov

Abstract

Data-driven deep learning (DL) methods using convolutional neural networks (CNNs)
demonstrate promising performance in natural image computer vision tasks. However,
using these models in medical computer vision tasks suffers from several limitations,
viz., (i) adapting to visual characteristics that are unlike natural images; (ii) modeling
random noise during training due to stochastic optimization and backpropagation-based
learning strategy; (iii) challenges in explaining DL black-box behavior to support
clinical decision-making; and (iv) inter-reader variability in the ground truth (GT)
annotations affecting learning and evaluation. This study proposes a systematic
approach to address these limitations for COVID-19 detection using chest X-rays
(CXRs). Specifically, our contribution benefits from (i) pretraining specific to CXRs in
transferring and fine-tuning the learned knowledge toward improving COVID-19
detection performance; (ii) using ensembles of the fine-tuned models to further improve
performance compared to individual constituent models; (iii) performing statistical
analyses at various learning stages to validate our claims; (iv) interpreting learned
individual and ensemble model behavior through class-selective relevance mapping
(CRM)-based region of interest (ROI) localization; (v) analyzing inter-reader variability
and ensemble localization performance using Simultaneous Truth and Performance
Level Estimation (STAPLE) methods. We observe that: (i) ensemble approaches
improved classification and localization performance; and, (ii) inter-reader variability
and performance level assessment helped guide algorithm design and parameter
optimization. To the best of our knowledge, this is the first study to construct
ensembles, perform ensemble-based disease ROI localization, and analyze inter-reader
variability and algorithm performance for COVID-19 detection in CXRs.

Introduction

Coronavirus disease 2019 (COVID-19) is caused by the new Severe Acute Respiratory
Syndrome Coronavirus 2 (SARS-CoV-2) that originated in Wuhan, China. The World
Health Organization (WHO) declared this disease spread as an ongoing pandemic [1].
As of July 6, 2020, the pandemic has resulted in over 11 million cases, and more than

530,000 deaths worldwide and continues to grow unabated. The disease commonly   6
infects the lungs and results in pneumonia-like symptoms [2]. Reverse   7
transcription-polymerase chain reaction (RT-PCR) analysis is the gold standard to   8
confirm infections. However, these tests are reported to exhibit varying sensitivity and   9
are not widely available [2]. Radiological imaging using chest X-rays (CXRs) and   10
computed tomography (CT) scans, though not currently recommended in the United   11
States, are commonly used radiological diagnostic support aids to manage COVID-19   12
disease progression [2]. While CT scans are more sensitive in detecting pulmonary   13
disease manifestations than CXRs, their use is limited due to issues such as   14
cross-contamination, non-portability, repeated sanitation requirements for CT   15
examination rooms, and equipment, and exposing patients under investigation (PUI),   16
hospital staff and technical personnel to the infection. Following the American College   17
of Radiology (ACR) recommendations [3], CXRs are considered a viable alternative to   18
CT scans in addressing some of these limitations. However, the pandemic nature of the   19
disease has compounded the existing shortage of expert radiologists, particularly in   20
third-world countries [4]. Under these circumstances, artificial intelligence (AI) driven   21
computer-aided diagnostic (CADx) tools have been considered as potentially viable   22
alternatives for facilitating swift patient referrals or aiding appropriate medical care [5].   23
Several studies using data-driven deep learning (DL) algorithms with convolutional   24
neural network (CNN) models in various strategies have been published for detecting,   25
localizing, or measuring progression of COVID-19 using CXRs and CTs [4] [6, 7]. While   26
there are scores of medical imaging CADx solutions that use DL approaches for disease   27
detection including COVID-19, there are significant limitations in existing approaches   28
related to data set type, size, scope, model architecture, and evaluation. We address   29
these shortcomings and propose novel analyses to meet the urgent demand for   30
COVID-19 detection using CXRs.   31

Image modality-specific transfer learning   32

Existing solutions tend to be disease-specific and require retraining on a   33
large-collection of expert-annotated data to ensure use in real-world applications.   34
Generalization of these approaches is challenged by available expert-annotations, their   35
strength (i.e. weak image-level labels versus strong region of interest (ROI) localizing   36
the pathology), and necessary computation resources. Under these circumstances,   37
transfer learning strategies are commonly adopted [8] where the models are trained on a   38
large-scale selection of stock photographic images like ImageNet [9] and then fine-tuned   39
for the specific task. A problem with this approach is that the architecture and   40
hyperparameters of these pre-trained models are optimized for natural image computer   41
vision applications. In contrast, medical image collections bearing the desired pathology   42
are significantly smaller in number. Therefore, using these models for medical visual   43
analyses often results in a covariate shift and generalization issues due to the difference   44
in source and target image modalities. Medical images are distinct in their   45
characteristics such as highly localized disease ROIs, and varying appearances for the   46
same disease label and severity [10]. Under these circumstances, the transferred   47
knowledge from the natural image processing domain, while seemingly successful in   48
disease classification, may not be optimal for disease localization. Medical images   49
exhibit different visual characteristics than natural images through high intra-class   50
variability and inter-class similarity, particularly for early-stage disease. To this end, we   51
propose training DL models with suitable depth on a large-scale selection of medical   52
images of the same modality to learn relevant feature representations that can be   53
transferred and fine-tuned for related medical visual recognition tasks. Such medical   54
modality-specific transfer learning could improve DL performance and generalization by   55
learning the common characteristics of the source and target modalities. This could lead   56
to a better initialization of model parameters and faster convergence, thereby reducing   57

computational demand, improving efficiency, and increasing opportunity for successful deployment. 58 59

Ensemble learning 60

Data-driven DL models use non-linear methods and learn through stochastic error backpropagation to perform automated feature extraction and classification. These models can only scale up in performance with an increase in the amount of training data and computational resources. Further, their sensitivity to the training data specifics limits their generalization due to learning different sets of weights at each instance of training. This stochastic learning nature results in different predictions referred to as the variance error. Also, there are issues concerning bias errors due to an oversimplified model that results in predictions that are different from the ground truth (GT) thereby placing a higher demand on appropriate threshold selection for obtaining desired performance. Ensemble learning seeks to address these issues by combining predictions of multiple models and resulting in a better performance compared to that of any individual constituent model [11]. There are several ensemble approaches such as majority voting, averaging, weighted averaging, stacking, blending, etc. 61 62 63 64 65 66 67 68 69 70 71 72 73

ROI localization 74

Data-driven medical DL models have often been maligned for their "black box" behavior, i.e., inability to make their decision-making process, which is critical for their clinical use, clear. This results in an apparent opaque relationship between input and predictions. This is often due to their massive architectural depth resulting in a large number of model parameters and lack of decomposability into individual explainable components. Further, multiple non-linear processing units perform complex data transformations that can result in unpredictable behavior. This opacity is a serious bottleneck in their use in deriving understandable clinical interpretations. 75 76 77 78 79 80 81 82

Variability in the ground truth (GT) 83

Supervised learning requires a consistent label associated with the appearance of the pathology in the image. However, in medical images, these labels can vary not only for disease stage and shared appearance with other diseases but also with observer expertise and sensitivity to assessment demands. A new pandemic, for example, may bias experts toward higher sensitivity, i.e. they will associate non-specific features with the new disorder because they lack experience with relevant disease manifestation in the image. Therefore, an assessment of observer variability constitutes an essential part of AI-based classification and localization studies. This includes analyzing (i) inter-reader, and (ii) intra-reader variability. It is reported that inter-reader variability tends to be higher than intra-reader variability because multiple observers may have a different opinion on the outlining disease-specific ROI depending on their expertise or personal leanings toward recommending necessary clinical care [12]. Thus, inter-reader variability is a serious bottleneck that may lead to misinterpretation through the "inexact" ROI annotations and also affects supervised learning. Not only can this lead to a false diagnosis or inability to evaluate the true benefit of accurately supplementing clinical-decision making, but it places a greater burden on the number of training images needed to overcome these implicit biases. Thus, it is imperative to conduct inter-reader variability analysis as part of evaluating AI performance. An obvious approach to overcome this challenge might be to compare a collection of annotations by several radiologists using relevant clinical data. However, quantifying expert performance in annotating disease-specific ROI is difficult. This persistent challenge exists because of the difficulty in obtaining or estimating a known true ROI for the task under study. While there exist automated tools to manage inter- and intra-reader variability, these algorithms need to be assessed to warrant their suitability for the task under study. Additionally, it is imperative to determine an appropriate measure for comparing individual expert annotations with each other and with the AI [13]. 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109

Lack of statistical analysis   110

Results and methods in a study need to be transparently reported to accurately   111
communicate scientific discovery. Statistical analyses are critical for measuring inherent   112
data variability and their impact on AI performance. They help in evaluating claims   113
and differentiating reasonable and uncertain conclusions. Statistical reporting helps to   114
alleviate issues resulting from incorrect data mining, biased samples, overgeneralization,   115
causality, and violating the assumptions concerning analysis. However, a study of the   116
literature reveals that scientific publications are often limited in presenting statistical   117
analyses of their results [14].   118

In this study, we address the aforementioned limitations through a stage-wise   119
systematic approach, as follows: (i) we explore the benefits of repeated CXR-specific   120
pretraining that results in learning CXR modality-specific knowledge, which can be   121
transferred and fine-tuned to improve performance toward COVID-19 detection in   122
CXRs; (ii) we compare the utility of several ImageNet-pretrained CNN models   123
truncated at their empirically determined intermediate layers to that of out-of-the-box   124
ImageNet-pretrained CNNs toward the current task; (iii) we use ensembles of fine-tuned   125
models for COVID-19 detection that are created through various strategies to improve   126
performance compared to any individual constituent model; (iv) we explain learned   127
behavior of individual CNNs and their ensembles using class-selective relevance   128
mapping (CRM)-based localization [15] tools that identify discriminative ROIs involved   129
in detecting COVID-19 viral disease manifestations; (v) we perform ensemble   130
localization to improve localization behavior and compensate for the error due to   131
neglected ROIs by individual CNNs; (vi) we perform exploratory studies to analyze   132
variability in model localization using annotations of two expert radiologists; (vii) we   133
measure statistical significance in performance metrics including Intersection over Union   134
(IoU) and mean average precision (mAP); and, (viii) we perform inter-reader variability   135
analysis using Simultaneous Truth and Performance Level Estimation (STAPLE) [13]   136
using a reference consensus annotation generated from the set of radiologists'   137
annotations. This is compared with individual radiologist annotations and the predicted   138
disease ROI by model ensembles to provide a measure of inter-reader variability and   139
algorithm performance. To our best knowledge, this is the first study to construct   140
ensembles, perform ensemble-based disease ROI localization, and evaluate inter-reader   141
reader variability and algorithm performance toward COVID-19 detection in CXRs.   142

Related Works   143

We describe related works for various topics discussed in this study below.   144

Image modality-specific transfer learning   145

The authors of [16] demonstrated the benefits of transferring knowledge learned from   146
training on a large-scale selection of CXR images and repurposing them toward   147
tuberculosis (TB) detection. They constructed model ensembles and compared their   148
performance with individual models toward classifying CXRs as showing normal lungs   149
or TB -like manifestations. The authors of [17] proposed CXR modality-specific   150
knowledge transfer by retraining the ImageNet-pretrained CNN models on a large-scale   151
selection of CXRs collected from various institutions. This helped in improving   152
generalization of the learned knowledge that was transferred and fine-tuned to detect   153
TB disease-like manifestations in CXRs. The authors performed ensemble learning   154
using the best-performing CNNs to demonstrate better performance in classifying CXRs   155
as belonging to normal or TB-infected classes. At present, the literature on CXR   156
analysis benefiting from modality-specific knowledge transfer particularly applied to   157
detect COVID-19 viral disease manifestations is limited. This leaves room for progress   158
toward evaluating the efficacy of these methods in improving the performance toward   159

COVID-19 detection.   160

Ensemble learning   161

The authors of [18] used model ensembles to classify CXRs as showing normal lungs   162
or TB-like radiological manifestations. It was observed that an ensemble of custom   163
CNN and ImageNet-pretrained models delivered superior classification performance   164
with an AUC of 0.99. The authors of [19] evaluated the efficacy of a stacked model   165
ensemble constructed from hand-crafted features/classifiers and DL models toward TB   166
detection in CXRs. CXRs collected from various institutions were used to improve the   167
generalization of the proposed approach. It was observed that the model ensembles   168
delivered better performance than individual constituent models in all performance   169
metrics. Ensemble learning has been applied to detect cardiomegaly in CXRs [20]. The   170
authors observed that DL model ensembles were 92% accurate as compared to 76.5%   171
accuracy obtained with hand-crafted features/classifiers. These results demonstrate the   172
superiority of ensemble learning over the traditional approach of evaluating the   173
performance with stand-alone models. Applied to COVID-19 detection in CXRs, the   174
authors of [5] iteratively pruned the DL models and constructed ensembles to improve   175
performance compared to individual constituent models. To this end, the authors   176
observed that the weighted average of iteratively pruned models demonstrated superior   177
classification performance with a 99.01% accuracy and AUC of 0.9972. Otherwise, the   178
literature available on applying ensemble learning toward COVID-19 detection in chest   179
radiographs is limited.   180

ROI localization   181

Exploratory studies in developing explainable and transparent AI solutions toward   182
clinical decision-making are crucial to developing robust solutions for clinical use.   183
Literature studies reveal several works interpreting the learned behavior of DL models   184
by highlighting pixels that impact prediction scores, with varying intensities. The   185
authors of [21] used deconvolution methods to modify the gradients that resulted in   186
qualitatively improving ROI localization. The authors of [22] inverted image   187
representations using up-CNN models to provide insights into learned feature   188
representations. The authors of [23] generated class-activation maps (CAM) by   189
mapping the prediction class scores back to the deepest convolutional layer. The   190
authors of [24] generalized the use of CAM tools and proposed gradient-weighted CAM   191
(Grad-CAM) methods that can be applied to CNNs with varying architecture. The   192
authors of [15] proposed the CRM algorithm to visualize discriminative ROIs in   193
classifying medical image modalities. The authors measured both positive and negative   194
contributions of the feature map spatial elements in the deepest convolutional layer of   195
the trained models toward making class-specific predictions. It was observed that CRM   196
methods delivered superior localization toward classifying medical imaging modalities   197
compared to CAM-based methods. Applied to the task of localizing COVID-19 viral   198
disease manifestations in CXRs and CT scans, the authors of [7] proposed a DL model   199
that learned the underlying feature representations from volumetric CT scans. It was   200
observed that the model showed better performance with an AUC of 0.96 in detecting   201
COVID-19 viral disease patterns and differentiating them from other non-COVID-19   202
pneumonia-related opacities. They used CAM-based visualization tools to localize the   203
suspicious ROIs toward detecting COVID-19 viral disease manifestations. The authors   204
of [25] proposed a custom DL model and used Grad-CAM tools to explain their   205
predictions toward COVID-19 detection. The model achieved a sensitivity of 83% in   206
detecting COVID-19 disease patterns in CXRs. The authors of [6] proposed a   207
weakly-labeled data augmentation approach to increase training data size for   208
recognizing COVID-19 viral related pneumonia opacities in CXRs. They used a   209
strategic approach to train various DL models with non-augmented and weakly-labeled   210
augmented training and evaluated their performance. It was observed that the simple   211

addition of CXRs showing COVID-19 viral disease manifestations to weakly labeled ₂₁₂
augmented training data improved performance. This study revealed that COVID-19 ₂₁₃
viral disease patterns have a uniquely different presentation compared to non-COVID-19 ₂₁₄
viral pneumonia-related opacities. The authors used Grad-CAM tools to study the ₂₁₅
behavior of models trained with non-augmented and augmented data toward localizing ₂₁₆
COVID-19 viral disease manifestations in CXRs. Otherwise, the literature is limited ₂₁₇
concerning the use of visualization tools toward COVID-19 detection in CXRs. ₂₁₈

Observer variability analysis ₂₁₉

Applied to CT scans, the authors of [26] analyzed inter- and intra-radiologist ₂₂₀
variability in detecting abnormal parenchymal lung manifestations on high-resolution ₂₂₁
CT scans. They used the Kappa statistic to measure the degree of agreement toward ₂₂₂
these analyses. A clinically acceptable agreement was observed between the radiologists, ₂₂₃
but the agreement rate declined when the radiologists were not involved in the regular ₂₂₄
analysis of thoracic CT scans. Another study [27] analyzed COVID-19 disease ₂₂₅
manifestations in high-resolution CT scans obtained from patients at the North Sichuan ₂₂₆
Medical College, Nanchong, China. They assessed inter-observer variability by having ₂₂₇
CT readers repeat the data analysis at intervals of three days. A comparison of a set of ₂₂₈
measurements by the same scan reader was used to assess intra-observer variability. ₂₂₉
They observed the existence of significant variability in inter- and intra-observer analysis, ₂₃₀
concerning the extent and density of disease spread. Applied to CXR analysis, the ₂₃₁
authors of [28] performed an observational study among Russian clinicians in analyzing ₂₃₂
the variability toward interpreting abnormalities in CXRs. The agreement was analyzed ₂₃₃
in different scales using the Kappa statistic for a set of 50 CXRs, using different scales. ₂₃₄
It was observed that there existed only a fair agreement in detecting and localizing ₂₃₅
abnormalities with a Kappa value of 0.380 and 0.448, respectively. This demonstrated ₂₃₆
that limited agreement on interpreting abnormalities resulted in sub-optimal population ₂₃₇
screening. At present, there is no available literature on the analysis of inter- and/or ₂₃₈
intra-reader variability applied to COVID-19 detection in CXRs. ₂₃₉

Statistical analysis ₂₄₀

The authors of [14] conducted a cross-sectional study toward analyzing the quality of ₂₄₁
statistical reporting in a random selection of publications in the Journal of Physiology ₂₄₂
and the British Journal of Pharmacology. The study used samples before and after the ₂₄₃
publication of an editorial, suggesting measures to adopt in reporting data and ₂₄₄
statistical analyses. The authors observed no evidence of change in reporting these ₂₄₅
measures after the editorial publication. It is observed that 90-96% of papers were not ₂₄₆
reporting statistical significance measures including p-values to identify the specific ₂₄₇
groups exhibiting these statistically significant differences in performance. Appropriate ₂₄₈
statistical analyses are included in the current study. ₂₄₉

Materials and methods ₂₅₀

Data collection ₂₅₁

This retrospective study uses the following publicly available datasets: ₂₅₂
i) Pediatric CXR dataset: The authors of [29] made available a collection of 5,856 ₂₅₃
pediatric CXRs showing normal lungs (n = 1,583or bacterial (n = 2,780) or viral ₂₅₄
pneumonia (n = 1,493) disease manifestations. The data were collected from children ₂₅₅
age 1 to 5 years at the Guangzhou Children's Medical Center, China. The radiological ₂₅₆
examinations were performed as a part of routine clinical care. The CXR images are ₂₅₇
made available in JPEG format, and approximately 2000 × 2000 pixels resolution with ₂₅₈
8-bit depth. ₂₅₉
ii) RSNA CXR dataset: The authors of [30] made available a collection of 26,684 ₂₆₀
frontal CXRs for a Kaggle challenge. The CXRs are grouped into t normal (n = 8,851) ₂₆₁

and abnormal (n = 17,833) classes; the abnormalities include pneumonia or non-pneumonia related opacities. The CXR images are made available in 1024 × 1024 8-bit pixels resolution and DICOM format.

iii) CheXpert CXR dataset: The authors of [31] made available a collection of 191,219 frontal CXRs showing normal lungs (n = 17,000) or other pulmonary abnormalities (n = 174,219). The CXR images are collected from patients at Stanford University Hospital, California, and are labeled for various thoracic disease manifestations by an automated natural language processing (NLP)-based labeler. The labels are extracted from radiological texts and conform to the Fleischner Society glossary of terms for thoracic imaging.

iv) NIH CXR-14 dataset: The authors of [8] released a collection of 112,120 frontal CXRs, collected from 30,805 patients at the NIH Clinical Center, Maryland. The collection includes CXRs, labeled as showing pulmonary abnormalities (n = 51,708) or normal lungs (n = 60,412). The CXRs were screened to remove personally identifiable information (PII) and ensure patient privacy. The CXRs belonging to the abnormal category are labeled for multiple thoracic disease manifestations using the information extracted from radiological reports using an automated NLP-based labeling algorithm.

v) Twitter-COVID-19 CXR dataset: A radiologist from a hospital in Spain made available a collection of 134CXRs exhibiting COVID-19 viral pneumonia manifestations, on Twitter (https://twitter.com/ChestImaging). The data were collected from SARS-CoV-2 PCR+ subjects and are made available at approximately 2000 ×2000 pixels resolution in JFIF format.

vi) Montreal-COVID-19 CXR dataset: The authors of [32] manage a GitHub repository that hosts a collection of CXRs and computed tomography (CT) scans of SARS-CoV-2 + and/or suspected patients. The images are pooled from publications and hospitals through collaboration with physicians and other public resources. As of May 20, 2020, the collection includes 226 CXRs showing COVID-19 viral pneumonia manifestations. The authors didn't provide complete metadata, however, the collection includes CXRs of 131 male patients and 64 female patients; the average age for the COVID-19 group is 58.8±14.9 years.

Lung ROI Cropping and preprocessing

Input data characteristics directly impact DL model learning, which is significant in applications that involve disease detection. For example, clinical decision-making could be adversely impacted by learning irrelevant features. In the case of COVID-19 and other pulmonary diseases, it is vital to limit analysis to the lung ROI and train the models toward learning relevant feature representations from within these pulmonary zones. Literature studies reveal that U-Net-based semantic segmentation delivers commendable performance in segmentation tasks using natural and medical imagery [33]. For this study, we are using a custom U-Net with dropout layers to segment the lung ROI from the background. Gaussian dropouts are used in the encoder, as shown in Fig 1, to reduce overfitting and provide restrictive regularization [34]. A dropout ratio of 0.5 is used after empirical pilot evaluations. The segmentation workflow is shown in Fig 2.

[Fig 1 about here.]

[Fig 2 about here.]

The model is trained on CXRs and their associated lung masks made available by the authors of [35]. Sigmoidal activation is used at the deepest convolutional layer to restrict the mask pixels to the range (0 − 1). The model is optimized to minimize a combination of binary cross-entropy and dice losses. Callbacks are used to store model checkpoints after each epoch. The best model weights are used for lung mask generation. The model is trained to generate lung masks at 256 × 256 pixel resolution

for various datasets used in this study. The lung boundaries are delineated using the ₃₁₂
generated masks and are cropped to a bounding box containing the lung pixels. The ₃₁₃
lung bounding boxes are resized to 256×256 pixel dimensions and used for further ₃₁₄
analysis. The cropped lung bounding boxes are further preprocessed as follows: (i) ₃₁₅
Images are normalized so that the pixel values are restricted to the range $(0 - 1)$. (ii) ₃₁₆
Images are passed through a median filter to perform noise removal and edge ₃₁₇
preservation. (iii) Image pixels are centered through mean subtraction and are ₃₁₈
standardized to reduce computational complexity. ₃₁₉

Repeated CXR-specific pretraining ₃₂₀

Fig 3 illustrates the workflow showing various stages of model training and ₃₂₁
evaluation. ₃₂₂

[Fig 3 about here.] ₃₂₃

First, the images are preprocessed to remove irrelevant features by cropping the lung ₃₂₄
ROI. The cropped images are used for model training and evaluation. We perform ₃₂₅
repeated CXR-specific pretraining in transferring modality-specific knowledge that is ₃₂₆
fine-tuned toward detecting COVID-19 viral manifestations in CXRs. Training proceeds ₃₂₇
in a series of steps. First, the CNNs are trained on a large collection of CXRs to separate ₃₂₈
normals from those showing abnormalities of any type. Next, we retrain the models ₃₂₉
from the previous step, focusing on separating CXRs showing bacterial pneumonia or ₃₃₀
non-COVID pneumonia from normals. Next, we fine-tune the models from the previous ₃₃₁
step toward the specific separation of CXRs showing COVID-19 pneumonia from ₃₃₂
normals. Finally, the learned features from these phases of training become parts of the ₃₃₃
ensembles developed to optimize the detection of COVID-19 pneumonitis from CXRs. ₃₃₄

The details of this step-wise approach are discussed as follows. In the first stage of ₃₃₅
pretraining, a custom CNN and selected ImageNet-pretrained CNN models are ₃₃₆
retrained on a large selection of CXRs with sufficient diversity due to sourcing from ₃₃₇
different collections, to coarsely learn the characteristics of normal and abnormal lungs. ₃₃₈
This CXR-specific pretraining helps in converting the weight layers, specific to the ₃₃₉
CXRs, in subsequent steps. Table 1 shows the distribution of data used in the first ₃₄₀
stage of repeated CXR-specific pretraining. ₃₄₁

Table 1. Data distribution for the first stage of repeated CXR-specific pretraining. A custom CNN and a selection of ImageNet-pretrained CNNs are retrained on a large selection of CXRs to learn CXR-specific features to categorize them as showing normal or abnormal lungs.

Dataset	Normal	Abnormal
RSNA	8331	17833
CheXpert	16480	17000
NIH	59892	51708
Total	84703	86541

The motivation behind this approach is to perform a knowledge transfer from the ₃₄₂
natural image domain to CXR-domain and learn the characteristics of normal lungs and ₃₄₃
a wide selection of CXR-specific pulmonary disease manifestations. During this training ₃₄₄
step, the datasets are split at the patient-level into 90% for training and 10% for testing. ₃₄₅
We randomly allocated 10% of the training data for validation. ₃₄₆

During the second stage of repeated CXR-specific pretraining, the learned knowledge ₃₄₇
from the first stage pretrained models is transferred and repurposed to classify CXRs as ₃₄₈
exhibiting normal lungs, bacterial pneumonia, or non-COVID-19 viral pneumonia ₃₄₉
manifestations. This pretraining is motivated by the biological similarity in ₃₅₀
non-COVID-19 viral and COVID-19 viral pneumonia. However, there exist distinct ₃₅₁

radiological manifestations between each other as well as with non-viral   352
pneumonia-related opacities [6] [29]. The motivation is to transfer the learned   353
knowledge and fine-tune for COVID-19 detection. Table 2 shows the datasets used and   354
their distribution for this pretraining stage. For the normal class, we pooled CXRs from   355
various collections to introduce generalization and improve model performance. During   356
this pretraining stage, again, the datasets are split at the patient-level into 90% for   357
training and 10% for testing. For validation, we randomly allocated 10% of the training   358
data.   359

Table 2. Data distribution for the second stage of repeated CXR-specific pretraining. The first-stage pretrained models are retrained on a collection of CXRs to categorize them as showing normal lungs, bacterial pneumonia, or non-COVID-19 viral pneumonia manifestations. Note that the pediatric CXR dataset predates the onset of the SARS-CoV2 virus, and therefore the viral pneumonia is of non-COVID-19 type.

Dataset	Normal	Bacterial pneumonia	Non-COVID-19-viral pneumonia
CheXpert	400	-	-
NIH	400	-	-
Pediatric CXR	1583	2780	1493
RSNA	400	-	-
Total	2783	2780	1493

Fine-tuning for COVID-19 detection   360

The learned knowledge from the second stage of pretraining is transferred and   361
fine-tuned to improve performance in classifying CXRs as showing normal lungs or   362
COVID-19 viral pneumonia disease manifestations. Table 3 shows the datasets used and   363
their distribution toward this fine-tuning stage. We compare this performance to that   364
without repeated CXR-specific pretraining, referred to as Baseline, where the   365
ImageNet-pretrained CNNs are retrained out-of-the-box to categorize the CXRs as   366
showing normal lungs or COVID-19 viral disease manifestations. For the normal class,   367
we pooled CXRs in a patient-specific manner from various collections to introduce   368
generalization and improve model performance. During this training step, we performed   369
a patient-level split of the train and test data as follows: The CXRs from the   370
Montreal-COVID-19 and Twitter-COVID-19 collections are combined (n = 360) where   371
n is the total number of images in the collection. The data is split at the patient-level   372
into 80% for training and 20% for testing. We randomly allocated 10% of the training   373
data for validation. The test set includes 72 CXRs, containing 36 CXRs each from the   374
Montreal-COVID-19 and Twitter-COVID-19 collections.   375

The GT disease annotations for this test data are set by the verification of publicly   376
identified cases from two expert radiologists, referred to as Rad-1 and Rad-2 hereafter,   377
with a combined experience of 60 years. The radiologists used the web-based VGG   378
Image Annotator tool [36] to independently annotate the COVID-19 viral   379
disease-specific ROI in the test collection. The radiologists were shown the chest   380
radiographs in Portable Network Graphics format with a spatial resolution of 1024 ×   381
1024 pixels and were asked to annotate COVID-19 viral disease-specific ROI in the   382
given test set.   383

Data augmentation   384

It is well known that large amounts of high-quality data are imperative for DL   385
model training and achieving superior performance. A challenge in the medical   386
image-based DL is the lack of sufficient data. Many studies limit their work to data   387
sourced from a single site. Using limited, single-site data toward model training may   388
result in loss of generalizability and degrade model performance when evaluated on   389
unseen data from other institutions or diverse imaging practices. Under these   390

Table 3. Data distribution for COVID-19 detection. The second-stage pretrained models are fine-tuned to classify CXRs into showing normal lungs or COVID-19 viral patterns.

Dataset	COVID-19+	Normal
CheXpert	-	120
Montreal-COVID-19	226	-
NIH	-	120
RSNA	-	120
Twitter-COVID-19	134	-
Total	360	360

circumstances, generalizability and performance could be improved by increasing the variability of training data. In this study, we use a diversified data distribution from multiple CXR collections to enhance model generalization and performance in repeated CXR-specific pretraining and fine-tuning stages. Class weights are used to reward the minority classes to prevent biasing error and reduce overfitting. During model training, data are augmented with random horizontal and vertical pixel shifts in the range (-5 –to 5) and rotations in the degree range (-9 –to 9). 391 392 393 394 395 396 397

Models 398

The following CNN-based DL models are trained and evaluated at various stages of learning performed in this study: (i) a custom wide-residual network (WRN) [37] with dropout, (ii) ResNet-18 [38], (iii) VGG-16 [39], (iv) VGG-19 [39], (v) Xception [40], (vi) Inception-V3 [41], (vii) DenseNet-121 [42], (viii) MobileNet-V2 [43], (ix) NasNet-Mobile [44]. The models are selected with an idea of increasing the architectural diversity, thereby increasing the representation power, when used in ensemble learning. All computation is done on a Windows® system with Intel Xeon CPU E3-1275 v6 3.80 GHz processor and NVIDIA GeForce® GTX 1050 Ti. We used Keras DL framework with Tensorflow backend, CUDA, and CUDNN libraries to accelerate GPU performance. 399 400 401 402 403 404 405 406 407

Residual CNNs having depths of hundreds of layers suffer from diminishing feature reuse [37]. This occurs due to issues with gradient flow, which results in only a few residual blocks learning useful feature representations. A WRN model combats diminishing feature reuse issues by reducing the number of layers and increasing model width [37]. The resultant networks are found to exhibit shorter training times with similar or improved accuracy. In this study, we use a custom WRN with dropout regularization. Dropouts provide restrictive regularization, address overfitting issues, and enhance generalization. After empirical observations, we used 5×5 kernels for the convolutional layers, assigned a dropout ratio of 0.3, a depth of 16, and a width of 4, for the custom WRN used in this study. The resultant architecture is referred further to as custom WRN. Fig 4 shows a WRN block with dropout used in this study. The output from the deepest residual block is average pooled, flattened, and appended to a final dense layer with Softmax activation to predict class probabilities. 408 409 410 411 412 413 414 415 416 417 418 419 420

[Fig 4 about here.] 421

As mentioned before, ImageNet-pretrained CNNs have been developed for computer vision tasks with natural images. These models have varying depth and learn diversified feature representations. For medical images that are often available in limited quantities, deeper models may not be optimal and can lead to overfitting and loss of generalization. During the first stage of pretraining, the CNNs are instantiated with 422 423 424 425 426

their ImageNet-pretrained weights and are truncated at empirically determined 427
intermediate layers to effectively learn the underlying feature representations for CXR 428
images and improve classification performance. The truncated models are appended 429
with (i) zero-padding, (i) a 3×3 convolutional layer with 1024 feature maps, (ii) a 430
global average pooling (GAP) layer, (iii) a dropout layer with an empirically determined 431
dropout ratio of 0.5, and (iv) a final dense layer with Softmax activation to output 432
prediction probabilities. These customized models learn CXR-specific feature 433
representations to classify CXR images into normal and abnormal. The custom WRN is 434
initialized with random weights. Fig 5 shows the architecture of the pretrained CNNs 435
used during the first stage of CXR-specific pretraining. 436

[Fig 5 about here.] 437

In the second stage, pretrained models from the first stage are truncated at their 438
deepest convolutional layer and appended with (i) GAP layer, (ii) dropout layer (ratio 439
= 0.5), and (iii) dense layer with Softmax activation to output class probabilities for 440
normal, bacterial pneumonia, and non-COVID-19 viral pneumonia. Fig 6 shows the 441
architecture of the customized models used during the second stage of pretraining. 442

[Fig 6 about here.] 443

Next, the second-stage pretrained models are truncated at their deepest 444
convolutional layer and appended with (i) GAP layer, (ii) dropout layer (ratio = 0.5), 445
and (iii) dense layer with Softmax activation. The resultant models are fine-tuned to 446
classify the CXRs as belonging to COVID-19+ or normal classes where '+' symbolizes 447
COVID-19-positive cases. Fig 7 shows the architecture of the models used toward 448
COVID-19 detection. The models in various learning stages are trained and evaluated 449
using stochastic gradient descent optimization to estimate learning error and 450
classification performance. We used callbacks to check the internal states of the models 451
and store model checkpoints. The model weights delivering superior performance with 452
the test data are used for further analysis. 453

[Fig 7 about here.] 454

The performance of the models at various learning stages is evaluated using the 455
following metrics: (i) Accuracy; (ii) Area under curve (AUC); (iii) Sensitivity; (iv) 456
Specificity; (v) Precision; (vi) F_1 score; (vii) Matthews correlation coefficient (MCC); 457
(viii) Kappa statistic; and (ix) Diagnostic Odds Ratio (DOR). 458

Ensemble Learning 459
The following ensemble strategies are applied to the fine-tuned models for COVID-19 460
detection to improve performance: (i) Majority voting; (ii) Simple averaging; and (iii) 461
Weighted averaging. In majority voting, the predictions with maximum votes are 462
considered as final predictions. The average of the individual model predictions is 463
considered the final prediction in a simple averaging ensemble. For a weighted ensemble, 464
we optimized the weights for the model predictions that minimized the total logarithmic 465
loss. This loss decreases as the prediction probabilities converge to true labels. We used 466
the Sequential Least Squares Programming (SLSQP) algorithmic method [45] to 467
perform several iterations of constrained logarithmic loss minimization to converge to 468
the optimal weights for the model predictions. 469

Inter-reader variability analysis 470
Fig 8 shows examples of COVID-19 viral disease-specific ROI annotations on CXRs 471
made by Rad-1 and Rad-2. In this study, we used the well-known STAPLE algorithm to 472
arrive at a consensus reference annotation and use it to evaluate the performance of the 473
top-N ensembles and to simultaneously assess the performance against each radiologist. 474

[Fig 8 about here.] 475

STAPLE methods are widely used in validating image segmentation algorithms and 476 comparing the performance of experts. Segmentation solutions are treated as a response 477 to a pixel-wise classification problem. The algorithm uses an expectation-maximization 478 (EM) approach that computes a probabilistic estimate of a reference segmented image 479 computed from a collection of expert annotations and weighing them by an estimated 480 level of performance for each expert. It incorporates this knowledge to spatially 481 distribute the segmented structures while satisfying homogeneity constraints. The 482 algorithm is summarized as follows: Let $Q = (q_1, q_2, \ldots, q_n)^N$ and $R = (r_1, r_2, \ldots,$ 483 $r_n)^N$ denote two column vectors, each containing A elements. The elements in Q and R 484 represent sensitivity and specificity parameters, respectively, characterizing one of N 485 segmentations. Let D denote an $M \times N$ matrix that describes segmentation decisions 486 made for each image pixel. Let N denote an indicator vector containing M elements 487 representing hidden, true binary segmentation values. The complete data can be written 488 as (D, N) and the probability mass function as f (D, N|q, r). The performance level of 489 the experts, characterized by a tuple (q, r) is estimated by the EM algorithm, which 490 maximizes (q', r'), the data log-likelihood function, given by, 491

$$(q', r') = argmax_{q,r} \ln(f(D, N|q, r)) \tag{1}$$

We used the following measures including Kappa statistic, sensitivity, specificity, 492 positive predictive value (PPV), and negative predictive value (NPV) to analyze 493 inter-reader variability and assess program performance. We used the 494 STAPLE-generated consensus ROI as to the standard reference and measured its 495 agreement with that generated by the top-N ensembles and the annotations of Rad-1 496 and Rad-2. We propose an algorithm to determine the set of True Positive (TP), False 497 Positive (FP), True Negative (TN), and False Negative (FN) for different IoU 498 thresholds in the range (0.1 – 0.7). The IoU evaluation metric, also called the Jaccard 499 Index, is widely used in object detection, given by a ratio as shown below: 500

$$IoU \; (Jaccard \; Index) = \frac{Area \; of \; overlap}{Area \; of \; union} \tag{2}$$

where Area of overlap measures the overlap between ROI annotations and Area of union 501 denotes their total combined area. An annotated ROI provided by a given radiologist or 502 that predicted by the top-N ensemble is considered as a TP if the IoU with the 503 STAPLE-generated consensus ROI is greater than or equal to a given IoU threshold. 504 Each radiologist or top-N ensemble predicted ROI that produces an IoU less than the 505 threshold or falls outside the consensus ROIs is counted as FP. The FN is defined as a 506 radiologist ROI or that predicted by the top-N ensemble that is completely missing 507 when there is an ROI in the consensus ROI. If there is an image with no ROIs on both 508 the masks under test, then we consider it as TN. The values are determined at 509 ROI-level per image and summed to calculate the Kappa statistic given by, 510

$$Kappa = 1 - \frac{1 - p_o}{1 - p_e} \tag{3}$$

where p_o is the measure of relative observed agreement and p_e denotes the agreement 511 through the hypothetical probability of chance. The values of p_o and p_e are computed 512 as follows: 513

$$p_o = \frac{(TP + TN)}{TP + FN + FP + TN} \tag{4}$$

$$p_e = \frac{p\_true}{p\_false} \tag{5}$$

$$p\_true = \frac{(TP + FN)(FP + TP)}{(TP + FN + FP + TN)^2} \tag{6}$$

$$p\_false = \frac{(FP + TN)(FN + TN)}{(TP + FN + FP + TN)^2} \tag{7}$$

The sensitivity, specificity, PPV, and NPV parameters are defined as, 514

$$Sensitivity = \frac{TP}{FN + TP} \tag{8}$$

$$Specificity = \frac{TN}{FP + TN} \tag{9}$$

$$PPV = \frac{TP}{FP + TP} \tag{10}$$

$$NPV = \frac{TN}{FN + TN} \tag{11}$$

Kappa values of 1 and 0 denote complete agreement and disagreement (other than 515
occurring by chance) among the readers, respectively. The value of Kappa becomes 516
negative if the agreement gets worse than random. The algorithm for measuring 517
inter-reader variability is given in Table 4, where m^1, m^2, and m^p denote the ROI 518
annotations of Rad-1, Rad-2, and that predicted by the top-N ensemble, respectively. 519

Disease ROI Localization 520

In this study, we use the CRM [15] visualization method to interpret the learned 521
behavior of individual models and their ensembles in localizing COVID-19 viral 522
disease-specific ROI manifestations. CRM has been shown to deliver better localization 523
performance than class-activation mapping (CAM)-based visualization. CRM-based 524
localization considers the fact that a feature map spatial element from the trained 525
model's deepest convolution layer would not only contribute to increasing the prediction 526
score for an expected class but also decreasing the score for other class outputs. This 527
helps in maximizing the gap between the scores for various classes. The process results 528
in highly-discriminative ROI localization since it uses incremental mean-squared error 529
(MSE) measured from the output nodes. We construct an ensemble of CRMs by 530
averaging those generated from various fine-tuned models for COVID-19 detection. The 531
size of CRMs from individual models is up-scaled to the size of the image input through 532
a normalization process. This is because the CRMs vary in size depending on the 533
feature map dimensions from the deepest convolutional layers of the individual models. 534
Based on empirical observations, the CRMs are thresholded to remove mapping scores 535
below 20% to alleviate noise resulting from low mapping scores when constructing CRM 536
ensembles. The resulting ensemble CRM localization is expected to compensate for the 537
error of missed ROI by individual models and enhance COVID-19 disease ROI 538
localization. 539

We evaluate the effectiveness of CRM-based ensemble localization through the 540
following steps. First, we use CRM-based ROI localization in interpreting the 541
predictions of individual CNNs and compare against the GT annotations provided by 542
each of the two experts. Next, we select the top-3, top-5, and top-7 performing models, 543

Table 4. Algorithm to assess inter-reader variability and program performance.

Algorithm

1: **Input:** Data $\left(m^1, m^2, m^p\right\}$, Threshold thr

2: **for** $i = 0, 1, 2, \ldots, N$ **do**

3: $m_i^{ref} = staple(m_i^1, m_i^2)$

4: **if** $(m_i^{ref} \, \mathbf{or} \, m_i^1/m_i^2/m_i^p)$ contains ROIs **then**

5: **for** ROI_j in $m_i^1/m_i^2/m_i^p$ **do**

6: **for** ROI_k in m_i^{ref} **do**

7: $metric = IOU(ROI_j, ROI_k)$

8: **if** $metric \geq thr$ **then**

9: $TP \leftarrow TP + 1$

10: **else if** $0 < metric < thr$ **or** missing ROI_k **then**

11: $FP \leftarrow FP + 1$

12: **else if** missing ROI_j **then**

13: $FN \leftarrow FN + 1$

14: **else**

15: $TN \leftarrow TN + 1$

16: **end for**

17: $p_0 = \frac{TP+TN}{TP+FP+FN+TN}$

18: $p_e = \frac{(TP+FN)(TP+FP)+(FP+TN)(FN+TN)}{(TP+FP+FN+TN)^2}$ 19: $Kappa = 1 - \frac{1-p_o}{1-p_e}$

20: $Sensitivity = \frac{TP}{FN+TP}$

21: $Specificity = \frac{TN}{FP+TN}$

22: $PPV = \frac{TP}{FP+TP}$

23: $NPV = \frac{TN}{FN+TN}$

24: **Output:** $Kappa$, $Sensitivity$, $Specificity$, PPV, and NPV

---

construct ensemble CRMs through an averaging process and compare against each 544
radiologists' independent annotations, and the STAPLE-generated consensus 545
annotation. Finally, we quantitatively compare the ensemble localization performance 546
with each other and against individual CRMs in terms of IoU and mean average 547
precision (mAP) metrics. The mAP score is calculated by taking the mean of average 548
precision (AP) over various IoU thresholds [46]. 549

Statistical Analysis 550

Statistical tests are conducted to determine significance in performance differences 551
between the models. We used confidence intervals (CI) to measure model discrimination 552
capability and estimate its precision through the error margin. We measured 95% CI as 553
the exact Clopper–Pearson interval for the AUC values obtained by the models in 554
various learning stages. Statistical packages including StatsModels and SciPy are used 555
in these analyses. We performed a one-way analysis of variance (ANOVA) [47] on mAP 556
values obtained with the top-N (N = (3, 5, 7)) ensemble to study their localization 557
performance and determine statistical significance among them and against the 558
annotations of each of the radiologist and also the STAPLE-generated consensus 559
annotation. One-way ANOVA tests are performed only if the assumptions of data 560
normality and homogeneity of variances are satisfied for which we performed 561
Shapiro-Wilk and Levene's analyses [47]. Statistical analyses are performed using R 562
statistical software (Version 3.6.1). 563

Results

Recall that in the first stage of CXR-specific pretraining, we truncated the
ImageNet-pretrained CNNs at their intermediate layers to empirically determine the
layers that demonstrated superior performance. These empirically determined layers for
the various models are shown in Table 5.

**Table 5. Candidate CNN layers delivering superior classification performance during the first stage of
CXR-specific pretraining.**

Model	Truncated layers
VGG-16	Block5-conv3
VGG-19	Block5-conv4
Inception-V3	Mixed3
Xception	Add_3
DenseNet-121	Pool3-pool
MobileNet-V2	Block_9_add
NASNet-mobile	Activation_94
ResNet-18	Add_6

The naming conventions for the layers are based on the Keras DL framework. The
performance achieved through truncating the models at the selected intermediate layers
and appending task-specific heads toward classifying the CXRs is shown in Table 6.

Table 6. Performance metrics achieved during the first-stage of CXR-specific pretraining. The custom WRN
is initialized with random weights. Data in parenthesis are 95% CI for the AUC values measured as the exact
Clopper–Pearson interval corresponding to separate 2-sided CI with individual coverage probabilities of $\sqrt{0.95}$. (Acc. =
Accuracy, AUC = Area under curve, Sens. = Sensitivity, Spec. = Specificity, Prec. = Precision, F_1 = F_1 score, MCC =
Matthews correlation coefficient, DOR = Diagnostics odd ratio). Bold numerical values denote best performances in the
respective columns. None of these individual differences are statistically significant.

Models	Acc.	AUC (CI)	Sens.	Spec.	Prec.	F_1	MCC	Kappa	DOR
Custom WRN	0.6696	0.722 (0.7153, 0.7287)	0.6566	0.6828	0.6763	0.6663	0.3395	0.3393	4.12
VGG-16	0.6874	0.7397 (0.7331, 0.7463)	0.6641	0.711	0.6988	0.6810	0.3755	0.3750	4.87
VGG-19	**0.6913**	**0.7435 (0.7374, 0.7506)**	0.6651	0.7178	0.704	0.6840	**0.3833**	**0.3827**	**5.06**
Inception-V3	0.6842	0.7375 (0.7309, 0.7441)	0.6186	0.7506	0.7145	0.6631	0.3723	0.3689	4.89
Xception	0.6727	0.7287 (0.7220, 0.7354)	0.6364	0.7094	0.6885	0.6614	0.3466	0.3456	4.28
DenseNet-121	0.6827	0.7416 (0.7350, 0.7482)	**0.7589**	0.606	0.6603	**0.7062**	0.3692	0.3650	4.85
NasNet-Mobile	0.6820	0.7347 (0.7281, 0.7413)	0.5802	**0.7849**	**0.7313**	0.6471	0.3728	0.3647	5.05
MobileNet-V2	0.6844	0.7426 (0.7360, 0.7492)	0.7007	0.668	0.6805	0.6904	0.3688	0.3686	4.72
ResNet-18	0.6821	0.7338 (0.7272, 0.7404)	0.7307	0.6332	0.6679	0.6979	0.3657	0.3640	4.69

From Table 6, we observe that the AUC values are not statistically significantly
different across the models ($p > 0.05$). The DOR provides a measure of diagnostic
accuracy and estimation of discriminative power. A high DOR is obtained by a model
that exhibits high sensitivity and specificity with low FPs and FNs. Considering AUC
and DOR values, VGG-19 demonstrates better performance followed by NasNet-Mobile
in classifying CXRs into normal and abnormal categories. A model with higher AUC
indicates that it is more capable of distinguishing TNs and TPs. Also considering MCC
and Kappa statistic metrics, VGG-19 outperformed other models. The confusion matrix,

ROC curves, and normalized Sankey flow diagram obtained using the VGG-19 model   580
toward this classification task are shown in Fig 9.   581

[Fig 9 about here.]   582

We used a normalized Sankey diagram [48] to visualize model performance. Here,   583
weights are assigned to the classes on the truth (left) and prediction (right) side of the   584
diagram to provide an equal visual representation for the classes on either side. The   585
strips width changes across the plot so that the width of each at the right side   586
represents the fraction of all objects which the model predicts as belonging to a category   587
that truly belongs to each of the categories.   588
Recall that during the second stage of CXR-specific pretraining, the learned   589
representations from the first-stage pretrained models are transferred and fine-tuned to   590
classify CXRs as showing normal lungs, bacterial proven pneumonia, or non-COVID-19   591
viral pneumonia. The performance achieved by the second-stage pretrained models is   592
shown in Table 7.   593

Table 7. Performance metrics achieved by the models during the second stage of CXR-specific pretraining. Bold numerical values denote best performances in the respective columns. None of these individual differences are statistically significant.

Models	Acc.	AUC (CI)	Sens.	Spec.	Prec.	F_1	MCC	Kappa	DOR
Custom WRN	0.7007	0.8589 (0.8332, 0.8846)	0.7007	0.8068	0.74	0.671	0.5326	0.5136	9.78
VGG-16	0.8879	0.9735 (0.9616, 0.9854)	0.8879	0.9298	0.896	0.8773	0.8312	0.8214	104.91
VGG-19	0.8922	0.9739 (0.9621, 0.9857)	0.8922	0.9304	0.906	0.8825	0.8389	0.8281	110.64
Inception-V3	0.9135	0.9792 (0.9699, 0.9895)	0.9135	0.9518	0.9120	0.9110	0.8656	0.8644	180.97
Xception	0.905	0.9714 (0.9590, 0.9838)	0.905	0.943	0.9064	0.9017	0.8532	0.8503	157.61
DenseNet-121	**0.9177**	**0.9835 (0.9740, 0.9930)**	**0.9177**	**0.9519**	**0.9187**	**0.9141**	**0.8736**	**0.8704**	**220.68**
NasNet-Mobile	0.9163	0.9819 (0.9720, 0.9918)	0.9163	0.9477	0.9222	0.9106	0.8674	0.8674	198.38
MobileNet-V2	0.9121	0.9812 (0.9711, 0.9913)	0.9121	0.952	0.9113	0.9098	0.8637	0.8621	205.81
ResNet-18	0.8936	0.9738 (0.9620, 0.9856)	0.8936	0.9329	0.8997	0.8849	0.8383	0.8309	116.77

We observed no statistically significant difference in AUC values achieved with the   594
models during this pretraining stage (p > 0.05). Considering DOR, DenseNet-121   595
demonstrated better performance (220.68) followed by MobileNet-V2 (205.81) in   596
categorizing the CXRs as showing normal lungs, bacterial pneumonia, or non-COVID-19   597
viral pneumonia. Considering MCC and F_1 score metrics that consider sensitivity and   598
precision to determine model generalization, DenseNet-121 outperformed other models.   599
The confusion matrix, ROC curves, and normalized Sankey flow diagram obtained using   600
the DenseNet-121 model toward this classification task are shown in Fig 10.   601

[Fig 10 about here.]   602

The second stage pretrained models are truncated at their deepest convolutional   603
layer, appended with task-specific heads, and fine-tuned to classify the CXRs as   604
belonging to COVID-19+ or normal categories. Table 8 shows the performance metrics   605
achieved by the models toward this task.   606
We observed no statistically significant difference in AUC values (p > 0.05) achieved   607
by the fine-tuned models. Considering DOR, ResNet-18 demonstrated better   608
performance (83.2) followed by DenseNet-121 (51.54) in categorizing the CXRs as   609
showing normal lungs or manifesting COVID-19 viral disease. The custom WRN,   610

Table 8. Performance metrics achieved with fine-tuning the second-stage pretrained models for COVID-19 detection. Bold numerical values denote best performances in the respective columns. Overall, ResNet-18 showed the best performance but individual metrics are not statistically different from other models.

Models	Acc.	AUC (CI)	Sens.	Spec.	Prec.	F_1	MCC	Kappa	DOR
D-WRN	0.8333	0.9043 (0.8562, 0.9524)	**0.9028**	0.7639	0.7927	0.8442	0.6732	0.6667	30.06
VGG-16	0.8681	0.9302 (0.8885, 0.9719)	0.8473	0.8889	0.8841	0.8653	0.7368	0.7361	44.4
VGG-19	0.8611	0.9176 (0.8726, 0.9626)	0.9028	0.8195	0.8334	0.8667	0.7248	0.7222	42.17
Inception-V3	0.8611	0.9123 (0.8660, 0.9586)	**0.9028**	0.8195	0.8334	0.8667	0.7248	0.7222	42.17
Xception	0.8681	0.9297 (0.8879, 0.9715)	0.8334	0.9028	0.8956	0.8634	0.7379	0.7361	46.47
DenseNet-121	0.875	0.9386 (0.8993, 0.9779)	**0.9028**	0.8473	0.8553	0.8784	0.7512	0.75	51.54
NasNet-Mobile	0.8542	0.911 (0.8644, 0.9576)	0.8612	0.8473	0.8494	0.8552	0.7085	0.7083	34.43
MobileNet-V2	0.875	0.925 (0.8819, 0.9681)	0.8473	0.9028	0.8971	0.8715	0.7512	0.75	51.54
ResNet-18	**0.8958**	**0.9490 (0.9132, 0.9854)**	0.8612	**0.9306**	**0.9254**	**0.8921**	**0.7936**	**0.7917**	**83.2**

Inception-V3, and DenseNet-121 are found to be equally sensitive (0.9028) toward this classification task. However, the ResNet-18 fine-tuned model demonstrated better performance with other performance metrics including accuracy, AUC, specificity, precision, F_1 score, MCC, and Kappa statistic. The confusion matrix, ROC curves, and normalized Sankey flow diagram obtained using the ResNet-18 model toward this classification task are shown in Fig 11.

[Fig 11 about here.]

Feature embedding visualization

We visualized the deepest convolutional layer feature embedding for the ResNet-18 fine-tuned model, using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [49]. We used t-SNE to visualize the embedding of the 1024-dimensional feature space into 2 dimensions, as shown in Fig 12. It is observed that the feature space for the normal and COVID-19+ classes is well-separated and clustered to facilitate the classification task.

[Fig 12 about here.]

The performance obtained with the fine-tuned models is compared to the Baseline, as shown in Table 9. The Baseline refers to out-of-the-box ImageNet-pretrained CNNs that are retrained toward this classification task. The custom WRN is initialized with randomized weights for the Baseline task.

As observed in Table 9, the fine-tuned models achieved better performance compared to their baseline counterparts. The AUC metrics achieved with the fine-tuned custom WRN, VGG-16, VGG-19, and NasNet-Mobile models are observed to be statistically significantly different ($p < 0.05$) compared to their baseline, untuned counterparts. We observed a marked reduction in the number of trainable parameters for the fine-tuned models. The fine-tuned DenseNet-121 model showed a 54.51% reduction in the number of trainable parameters while delivering better performance as compared to its baseline counterpart. The same holds true for ResNet-18 (46.05%), Inception-V3 (42.36%), Xception (37.57%), MobileNet-V2 (37.38%), and NasNet-Mobile (11.85%) with the added benefit of improved performance compared to their baseline models.

ROI visualization

We performed visualization studies to compare how the fine-tuned models and their baseline counterparts localize the ROIs in a CXR manifesting COVID-19 viral patterns.

Table 9. Performance metrics achieved during fine-tuning the second-stage pretrained models for COVID-19 detection is compared with theBaseline . The Baseline refers to retraining out-of-the-box ImageNet-pretrained CNNs toward this task. Bold numerical values show a reduction in the number of parameters.

Models	Method	Acc.	AUC (CI)	Sens.	Spec.	Prec.	F$_1$	MCC	Kappa	DOR	Para. Reduction (%)
Custom WRN	Baseline	0.7897	0.8014 (0.7362, 0.8666)	0.6742	0.8675	0.8396	0.7478	0.5611	0.5433	14.34	-
	Fine-tuned	0.8333	0.9043 (0.8562, 0.9524)	0.9028	0.7639	0.7927	0.8442	0.6732	0.6667	30.06	0
VGG-16	Baseline	0.7708	0.7993 (0.7338, 0.8648)	0.6667	0.875	0.8422	0.7442	0.5539	0.5416	14.01	-
	Fine-tuned	0.8681	0.9302 (0.8885, 0.9719)	0.8473	0.8889	0.8841	0.8653	0.7368	0.7361	44.4	0
VGG-19	Baseline	0.7847	0.8176 (0.7545, 0.8807)	0.8334	0.7362	0.7595	0.7948	0.5722	0.5694	13.97	-
	Fine-tuned	0.8611	0.9176 (0.8726, 0.9626)	0.9028	0.8195	0.8334	0.8667	0.7248	0.7222	42.17	0
Inception-V3	Baseline	0.8472	0.9285 (0.8864, 0.9706)	0.8473	0.8473	0.8473	0.8473	0.6945	0.6944	30.79	-
	Fine-tuned	0.8611	0.9123 (0.8660, 0.9586)	0.9028	0.8195	0.8334	0.8667	0.7248	0.7222	42.17	**42.36**
Xception	Baseline	0.8472	0.9215 (0.8775, 0.9655)	0.9028	0.7917	0.8125	0.8553	0.6988	0.6944	35.31	-
	Fine-tuned	0.8681	0.9297 (0.8879, 0.9715)	0.8334	0.9028	0.8956	0.8634	0.7379	0.7361	46.47	**37.57**
DenseNet-121	Baseline	0.8333	0.9153 (0.8698, 0.9608)	0.9028	0.7639	0.7927	0.8442	0.6732	0.6667	30.06	-
	Fine-tuned	0.8750	0.9386 (0.8993, 0.9779)	0.9028	0.8473	0.8553	0.8784	0.7512	0.75	51.54	**54.51**
NasNet-Mobile	Baseline	0.7778	0.8502 (0.7919, 0.9085)	0.8473	0.7084	0.744	0.7923	0.561	0.5556	13.48	-
	Fine-tuned	0.8542	0.911 (0.8644, 0.9576)	0.8612	0.8473	0.8494	0.8552	0.7085	0.7083	34.43	**11.85**
MobileNet-V2	Baseline	0.8681	0.9325 (0.8915, 0.9735)	0.8473	0.8889	0.8841	0.8653	0.7368	0.7361	44.4	-
	Fine-tuned	0.8750	0.925 (0.8819, 0.9681)	0.8473	0.9028	0.8971	0.8715	0.7512	0.75	51.54	**37.38**
ResNet-18	Baseline	0.8542	0.9302 (0.8885, 0.9719)	0.9167	0.7917	0.8149	0.8628	0.714	0.7083	41.83	-
	Fine-tuned	0.8958	0.9477 (0.9130, 0.9850)	0.8612	0.9306	0.9254	0.8921	0.7936	0.7917	83.2	**46.05**

Fig 13 shows the following: (i) a CXR with COVID-19 disease consensus ROI obtained with STAPLE using Rad-1 and Rad-2 annotations, and (ii) the ROI localization achieved with various fine-tuned models and their baseline counterparts. 643 644 645

[Fig 13 about here.] 646

We extracted the features from the deepest convolution layer of the fine-tuned models and their baseline counterparts. We used CRM tools to localize the pixels involved in predicting the CXR images as showing COVID-19 viral disease patterns. As 647 648 649

observed in Fig. 13, the baseline models demonstrated poor disease ROI localization, ₆₅₀
compared to the fine-tuned models. We observed that the fine-tuned models learned ₆₅₁
salient ROI feature representations, matching the experts' knowledge about the disease ₆₅₂
ROI. The localization excellence of the fine-tuned models can be attributed to (i) ₆₅₃
CXR-specific knowledge transfer that helped to learn modality-specific characteristics; ₆₅₄
the learned feature representations are transferred and repurposed for the COVID-19 ₆₅₅
detection task, and (ii) optimal architecture depth to learn the salient ROI feature ₆₅₆
representations to classify CXRs to their respective categories. These deductions are ₆₅₇
supported by poor localization performance of deeper, out-of-the-box ₆₅₈
ImageNet-pretrained baseline CNNs like DenseNet-121, Inception-V3, and ₆₅₉
MobileNet-V2, which possibly suffered from overfitting and resulted in poor learning ₆₆₀
and generalization. ₆₆₁

Ensemble studies ₆₆₂

We constructed ensembles of the top-3, top-5, and top-7 performing fine-tuned ₆₆₃
CNNs to evaluate for an improvement in predicting the CXRs as showing normal lungs ₆₆₄
or COVID-19 viral disease patterns. We used majority voting, simple averaging, and ₆₆₅
weighted averaging strategies toward this task. In weighted averaging, we optimized the ₆₆₆
weights for the model predictions to minimize the total logarithmic loss. We used the ₆₆₇
SLSQP algorithm to iterate through this minimization process and converge to the ₆₆₈
optimal weights for the model predictions. The results achieved with the various ₆₆₉
ensemble methods are shown in Table 10. ₆₇₀

Table 10. Performance achieved with an ensemble of top-3, top-5, and top-7 fine-tuned models toward COVID-19 detection. Bold numerical values denote best performances in the respective columns. Top-3 weighted averaging looks best but the AUC differences are not statistically significant.

Ensemble method	Top-N models	Acc.	AUC (CI)	Sens.	Spec.	Prec.	F_1	MCC	Kappa	DOR
Majority voting	3	0.9028	0.9097 (0.8628, 0.9566)	0.8612	0.9167	0.9155	0.8986	0.8084	0.8055	102.22
	5	0.8819	0.8819 (0.8291, 0.9347)	0.8612	0.9028	0.8986	0.8795	0.7646	0.7639	57.63
	7	0.8889	0.8889 (0.8375, 0.9403)	0.875	0.9028	0.9000	0.8874	0.7781	0.7778	65.02
Simple averaging	3	0.8958	0.9483 (0.9121, 0.9845)	0.8889	0.9028	0.9015	0.8952	0.7918	0.7917	74.32
	5	0.8819	0.9462 (0.9093, 0.9831)	0.8612	0.9028	0.8986	0.8795	0.7646	0.7639	57.63
	7	0.8819	0.9453 (0.9081, 0.9825)	0.875	0.8889	0.8874	0.8812	0.764	0.7639	56.01
Weighted averaging	**3**	**0.9097**	**0.9508 (0.9118, 0.9844)**	**0.9028**	**0.9445**	**0.9394**	**0.9091**	**0.8196**	**0.8194**	**105.6**
	5	0.9028	0.9493 (0.9134, 0.9852)	0.875	0.9306	0.9265	0.9000	0.8069	0.8055	93.87
	7	0.8889	0.9459 (0.9089, 0.9829)	0.8889	0.8889	0.8889	0.8889	0.7778	0.7778	64.02

We observed no statistically significant difference in the AUC values achieved by the ₆₇₁
various ensemble methods (p > 0.05). We observed that the performance with top-3 ₆₇₂
ensembles is better than that of top-5 and top-7 ensembles. It is observed that the ₆₇₃
weighted averaging of top-3 fine-tuned CNNs viz. ResNet-18, MobileNet-V2, and ₆₇₄
DenseNet-121 demonstrated better performance when their predictions are optimally ₆₇₅
weighted at 0.6357, 0.1428, and 0.2216, respectively. This weighted averaging ensemble ₆₇₆

delivered better performance in terms of accuracy, AUC, DOR, Kappa, F_1 score, MCC, and other metrics, as compared to other ensembles. The confusion matrix, ROC curves, and normalized Sankey flow diagram obtained with the weighted averaging of the top-3 fine-tuned CNNs are shown in Fig 14. Table 11 shows the performance achieved in terms of CRM-based IoU and mAP scores by the individual fine-tuned CNNs using the annotations of Rad-1, Rad-2, and STAPLE-generated consensus ROI.

[Fig 14 about here.]

Table 11. Performance achieved in terms of CRM-based IoU and mAP values by the individual fine-tuned CNNs using the radiologists' annotations and STAPLE-generated ROI consensus annotation. Bold numerical values denote best performances in the respective rows.

Annota-tions	Parame-ters	Xcep-tion	Inception-V3	DenseNet-121	VGG-19	VGG-16	MobileNet-V2	ResNet-18	NasNet-Mobile
Rad-1	IOU	0.0678	**0.1174**	0.0799	0.0854	0.1076	0.0644	0.0972	0.1000
	mAP @[0.1:0.7]	0.0571	**0.1142**	0.0697	0.0645	0.0986	0.0712	0.0593	0.075
	Ranking	8	**1**	5	6	2	4	7	3
Rad-2	IOU	0.2146	0.2567	0.2398	0.2183	0.2230	0.1825	0.2293	**0.2569**
	mAP @[0.1:0.7]	0.146	0.206	0.1858	0.1643	0.1882	0.1467	0.1742	**0.2186**
	Ranking	8	2	4	6	3	7	5	**1**
STAPLE	IOU	0.0670	**0.1337**	0.0916	0.0951	0.1267	0.0713	0.1126	0.1095
	mAP @[0.1:0.7]	0.0603	**0.1213**	0.0792	0.073	0.1068	0.0775	0.0648	0.0851
	Ranking	8	**1**	4	6	2	5	7	3

We observed that the model ROI predictions achieved varying IoU and mAP scores with the annotations of Rad-1, Rad-2, and the STAPLE-generated ROI consensus. For Rad-1, the fine-tuned Inception-V3 model demonstrated higher values for the average IoU and mAP metrics. For Rad-2, we observed that the fine-tuned NasNet-Mobile outperformed other models. With STAPLE-generated consensus ROI, the Inception-V3 model outperformed other models in localizing COVID-19 viral disease-specific ROI.

The precision-recall (PR) curves of the best performing models using Rad-1, Rad-2, and the STAPLE-generated consensus ROI are shown in Fig 15. These curves are generated for varying IoU thresholds in the range $(0.1 - 0.7)$. The confidence score threshold is varied to generate each curve. For a given fine-tuned model, we define the confidence score as the highest heat map value in the predicted ROI weighted by the classification score at the output nodes. We considered the ROI predictions as TP when the IoU and confidence scores are higher than their corresponding thresholds. For a given PR curve, we computed the AP score as the average of the precision across all recall values.

[Fig 15 about here.]

The following are the important observations from this localization study: (i) The accuracy of a model is not related to disease ROI localization. From Table 6, we observed that the fine-tuned ResNet-18 model is highly accurate, followed by DenseNet-121 and MobileNet-V2, in classifying the CXRs as belonging to the COVID-19 viral category. However, while localizing disease-specific ROI, the Inception-V3, VGG-16, and NasNet-Mobile fine-tuned models delivered superior ROI

localization performance compared to other models. This underscores the fact that the ⁷⁰⁶
classification accuracy of a model is not an optimal measure to interpret their learned ⁷⁰⁷
behavior. Localization studies are indispensable to understand the learned features and ⁷⁰⁸
compare them to the expert knowledge for the problem under study. These studies ⁷⁰⁹
provide comprehensive qualitative and quantitative measures of the learning capacity of ⁷¹⁰
the model and its generalization ability. ⁷¹¹

Next, we constructed an ensemble of CRMs through averaging the ROI localization ⁷¹²
for the top-3, top-5, and top-7 fine-tuned models. We ranked the models based on the ⁷¹³
IoU and mAP scores. The localization performance achieved with the various ensemble ⁷¹⁴
CRMs is shown in Table 12. ⁷¹⁵

Table 12. IOU and mAP values obtained by top-3, top-5, and top-7 ensembles using annotations of Rad-1, Rad-2, and STAPLE-generated consensus ROI annotations. Bold numerical values denote best performances in the respective rows.

Annotations	Parameters	Top-3	Top-5	Top-7
Rad-1	IOU	**0.1343**	0.0994	0.1236
	mAP@[0.1:0.7]	**0.1264**	0.0767	0.0753
Rad-2	IOU	0.2673	**0.2955**	0.2865
	mAP@[0.1:0.7]	0.2179	**0.2352**	0.2292
STAPLE	IOU	**0.1518**	0.1193	0.1350
	mAP@[0.1:0.7]	**0.1352**	0.0924	0.0916

From Table 12, we observed that the ensemble CRMs delivered superior ROI ⁷¹⁶
localization performance compared to that achieved with the individual models. ⁷¹⁷
However, the number of models in the top-performing ensembles varied. While using ⁷¹⁸
the annotations of Rad-1, we observed that the ensemble of the top-3 models ⁷¹⁹
demonstrated higher values for IoU and mAP than other ensembles. However, for ⁷²⁰
Rad-2, the ensemble of the top-5 models demonstrated superior localization with IoU ⁷²¹
and mAP values of 0.2955 and 0.2352, respectively. The ensemble of top-3 fine-tuned ⁷²²
models demonstrated higher values for IoU and mAP scores compared to other models ⁷²³
while using STAPLE-generated ROI consensus annotation. Considering this study, we ⁷²⁴
observed that averaging the CRMs of more than top-5 fine-tuned models didn't improve ⁷²⁵
performance but rather it saturates ROI localization. The PR curves obtained with the ⁷²⁶
top-N ensemble CRMs using Rad-1, Rad-2, and STAPLE-generated consensus ROI are ⁷²⁷
shown in Fig 16. ⁷²⁸

[Fig 16 about here.] ⁷²⁹

Instances of CXRs showing ROI annotations of Rad-1, Rad-2, top-3 ensemble using ⁷³⁰
STAPLE-generated ROI consensus (referred to as program hereafter), and the ⁷³¹
STAPLE-generated ROI consensus annotation are shown in Fig 17. ⁷³²

[Fig 17 about here.] ⁷³³

Fig 18 shows the following: (A) an ensemble CRM generated with the top-3 ⁷³⁴
fine-tuned models that delivered superior localization performance using ⁷³⁵
STAPLE-generated ROI consensus annotation, and (B) an ensemble CRM generated ⁷³⁶
with the top-5 fine-tuned models that delivered superior localization performance using ⁷³⁷
the annotations of Rad-2. ⁷³⁸

[Fig 18 about here.] ⁷³⁹

We observe that the CRMs obtained using individual models in the top-N ensemble ₇₄₀ highlight ROI to varying extents. The ensemble CRM averages the ROIs localized with ₇₄₁ individual CRMs to highlight the disease-specific ROI involved in class prediction. The ₇₄₂ ensemble CRMs have a superior IoU value, compared to that of individual CRMs; the ₇₄₃ ensemble CRM improved localization performance as compared to individual ROI ₇₄₄ localization. This underscores the fact that ensemble localization improves performance ₇₄₅ and ability to generalize, conforming to the experts' knowledge about COVID-19 viral ₇₄₆ disease manifestations. ₇₄₇

Statistical Analysis ₇₄₈

To perform a one-way ANOVA analysis, we investigated whether the assumptions of ₇₄₉ data normality and homogeneous variances are satisfied. We used the Shapiro–Wilk test ₇₅₀ to investigate for normal distribution of the data and Levene's test, for homogeneity of ₇₅₁ variances, using mAP scores obtained with the top-N ensembles. We plotted the ₇₅₂ residuals to investigate if the assumption of normal residual distribution is satisfied. Fig ₇₅₃ 19 shows the following: (A) The mean plot for the mAP scores obtained by the top-N ₇₅₄ ensembles using Rad-1, Rad-2, and STAPLE-generated consensus ROI annotations, and ₇₅₅ (B) a plot of the quantiles of the residuals against that of the normal distribution. ₇₅₆

[Fig 19 about here.] ₇₅₇

It is observed from Fig 19B that all the points fall approximately along with a ₇₅₈ 45-degree reference that shows that the assumption of the normal distribution is ₇₅₉ satisfied. Table 13 shows the consolidated results of Shapiro–Wilk, Levene, and one-way ₇₆₀ ANOVA analyses. ₇₆₁

Table 13. Consolidated results of Shapiro–Wilk, Levene, and one-way ANOVA analyses.

Metric	Shapiro–Wilk test (p)	Levene's test (p)	ANOVA (F)	ANOVA (p)
mAP	0.1014	0.3365	1.678	0.2060

To compute one-way ANOVA, we measure the variance between group means, the ₇₆₂ variance within the group, and the group sizes. This information is combined to ₇₆₃ measure statistical significance from the test statistic F if it follows an F-distribution. ₇₆₄ In our study, we have three groups (Rad-1, Rad-2, and STAPLE) of 10 observations ₇₆₅ each, hence the distribution is mentioned as F (2, 27). As observed from Table 13, the ₇₆₆ p-values obtained with the Shapiro-Wilk test are not significant ($p > 0.05$) and reveals ₇₆₇ that the normality assumption is satisfied. The result of Levene's test is not statistically ₇₆₈ significant ($p > 0.05$). That demonstrates that the variance across the mAP values ₇₆₉ obtained with the annotations of Rad-1, Rad-2, and STAPLE-generated consensus ROI ₇₇₀ are not statistically significantly different. Since the conditions of data normality and ₇₇₁ homogeneity of variances are satisfied, we performed one-way ANOVA to explore the ₇₇₂ existence of a statistically significant difference in the mAP scores. To this end, we ₇₇₃ observed no statistically significant difference in the mAP scores obtained with Rad-1, ₇₇₄ Rad-2, and STAPLE-generated consensus ROI (F (2, 27) = 1.678, p = 0.2060). This ₇₇₅ smaller F-value underscores the fact that the null hypothesis (H_0) -all groups ₇₇₆ demonstrate equal mAP scores- holds good. ₇₇₇

Inter-reader variability analysis and performance assessment ₇₇₈

We used the STAPLE-generated consensus ROI as to the standard reference and ₇₇₉ measured its agreement with that generated by the program and the radiologists. The ₇₈₀ consensus ROI is estimated from the set of ROI annotations provided by Rad-1 and ₇₈₁ Rad-2. STAPLE assumes that Rad-1 and Rad-2 individually annotated ROIs for the ₇₈₂ given CXRs so that the quality of annotations are captured. We determined the set of ₇₈₃ TPs, FPs, TNs, and FNs for 10 different IoU thresholds in the range (0.1 − 0.7) and ₇₈₄

provided a measure of inter-reader variability and program performance using the ₇₈₅
following metrics: (i) Kappa statistic; (ii) Sensitivity; (iii) Specificity; (iv) PPV; and (v) ₇₈₆
NPV. These parameters depend on the relative proportion of the disease-specific ROI. ₇₈₇
An ROI provided by a radiologist or predicted by the program is considered as a TP if ₇₈₈
the IoU with the consensus ROI is greater than or equal to a given IoU threshold. Each ₇₈₉
radiologist or program ROI that produces an IoU less than the threshold or falls outside ₇₉₀
the consensus ROIs is counted as FP. The FN is defined as a radiologist or program ₇₉₁
ROI that is completely missing when there is a consensus ROI. If there is an image with ₇₉₂
no ROIs on both the ROI annotations under test, it is considered as TN. Fig 20 shows ₇₉₃
the variability in Kappa, sensitivity, specificity, and PPV values observed for the Rad-1, ₇₉₄
Rad-2, and the program. ₇₉₅

[Fig 20 about here.] ₇₉₆

The estimated Kappa, sensitivity, specificity, PPV, and NPV values that are ₇₉₇
averaged over 10 different IoU thresholds in the range $(0.1 - 0.7)$ are shown in Table 14. ₇₉₈

Table 14. Performance level assessment and inter-reader variability analysis using STAPLE-generated consensus ROI. Bold numerical values denote the best performances in respective columns.

Annotations	Kappa	Sensitivity	Specificity	PPV	NPV
Rad - 1	**0.1805**	**1.0**	0.1384	**0.7140**	**1.0**
Rad - 2	0.0080	**1.0**	0.0121	0.2877	**1.0**
Program	0.0740	0.9037	**0.1467**	0.5154	0.6

The performance assessment as observed from Table 14 indicated that Rad-1 is more ₇₉₉
specific than Rad-2. The same holds good for the Kappa and PPV metrics. We ₈₀₀
observed that NPV is 1 for Rad-1 and Rad-2. This is because the number of FNs = 0, ₈₀₁
which signifies that none of the radiologists ROI completely missed when there is an ₈₀₂
ROI in the STAPLE-generated consensus annotation. However, the NPV achieved with ₈₀₃
the program is 0.6 which underscores the fact the predicted ROIs missed a marked ₈₀₄
proportion of ROIs in the STAPLE-generated consensus. This assessment indicated that ₈₀₅
Rad-1 generated annotations similar to that of STAPLE-generated consensus by ₈₀₆
demonstrating higher values for Kappa, sensitivity, and PPV as compared to Rad-2. We ₈₀₇
also observed that the program is performing with higher specificity but with lower ₈₀₈
sensitivity as compared to Rad-1 and Rad-2. These assessments provided feedback ₈₀₉
indicating the need for program modifications, parameter tuning, and other measures, ₈₁₀
to improve its localization performance. ₈₁₁

Discussion ₈₁₂

There are several salient observations to be made from the analyses reported above. ₈₁₃
These include the kind of data used in training, the size and variety of data collections, ₈₁₄
learning ability of various DL architectures informing their selection, need for ₈₁₅
customizing the models for improved performance, benefits of ensemble learning, and ₈₁₆
the imperative for localization to measure conformity to the problem. ₈₁₇
We observed that repeated CXR-specific pretraining and fine-tuning resulted in ₈₁₈
improved performance toward COVID-19 detection as compared to the baseline, ₈₁₉
out-of-the-box, ImageNet pretrained CNNs. This highlights the need to use task-specific ₈₂₀
modality training resulting in improved model adaption, convergence, reduced bias, and ₈₂₁
reduced overfitting. This approach may have helped the DL models differentiate ₈₂₂
distinct radiological manifestations between COVID viral pneumonia and other ₈₂₃

non-viral pneumonia-related opacities. An added benefit is that this approach resulted   824
in reductions in both computations and the number of trainable parameters.   825

It is well-known that neural networks develop or learn implicit rules to convert input   826
data into features for making decisions. These learned rules are opaque to the user and   827
the decisions are difficult to interpret. However, an interpretable model explaining its   828
predictions related to model accuracy doesn't necessarily guarantee that accurate   829
predictions are for the right reasons. Localization studies help observe if the model has   830
learned salient ROI feature representations that agree with expert annotations. In our   831
study, we demonstrate that CRM visualization tools show superior localization   832
performance in localizing COVID-19 viral disease-specific ROIs, particularly for the   833
fine-tuned models compared to the ImageNet-pretrained CNNs.   834

Model ensembles further improved qualitative and quantitative performance in   835
COVID-19 detection. Ensemble learning compensated mislabeling in individual models   836
by combining their predictions and reduced prediction variance to the training data.   837
Sensitivity also declined slightly, but this decline was not statistically significant. We   838
observed that the weighted averaging ensemble of the top-3 performing fine-tuned   839
models delivered better performance compared to any individual constituent model.   840
The results demonstrate that the detection task benefits from an ensemble of repeated   841
CXR-specific pretrained and fine-tuned models. Ensemble learning also compensates for   842
localization errors in CRMs and missed ROIs by combining and averaging the individual   843
CRMs. Empirical evaluations show that ensemble localization demonstrated superior   844
IoU and mAP scores and they significantly outperform ROI localization by individual   845
CNN models.   846

It is difficult to quantify individual radiologists' performance in annotating ROIs in   847
medical images. Not only are they the truth standard, but this "truth" is impacted by   848
inherent biases related to a pandemic event like COVID-19 and their clinical exposure   849
and experience. This complexity is compounded further because CXRs offer lower   850
diagnostic sensitivity than CTs for example. So, a conservative assessment of the CXR   851
is likely to result in smaller and more specific truth annotation ROIs. We used STAPLE   852
to compute a probabilistic estimate of expert ROI annotations for the two expert   853
radiologists who contributed to this study. STAPLE assumes these annotations are   854
conditionally independent. The algorithm discovers and quantifies the bias among the   855
experts when they differ in their opinion of the disease-specific ROI annotation. We use   856
STAPLE-generated annotations as GT to assess the variation for every annotation for   857
each expert, where the DL model is also considered as an expert. We observed that the   858
Kappa values obtained using the STAPLE-generated consensus ROI are in a low range   859
(0 – 0.2). This is probably because of the small number of experts and their inherent   860
biases in assessing COVID-19 cases. Particularly, we note that Rad-1 was very specific   861
in marking the ROIs, whereas Rad-2 annotated larger regions that sometimes   862
accommodated multiple smaller regions into a single ROI. This led to lower IoU value   863
that in turn affected the Kappa value. The pandemic is an evolving situation and CXR   864
manifestations often exhibit biological similarity to non-COVID-19 viral pneumonia.   865
The CXR is not a definitive diagnostic tool and expert views may differ in referring a   866
candidate patient for further review. It would be helpful to conduct a similar analysis   867
with a larger number of experts on a larger patient population. We remain hopeful that   868
health agencies and medical societies will make such image collections available for   869
future research. As more reliable and widely available COVID testing becomes available,   870
the results of that testing could be used with CXRs as an additional important   871
indicator of GT.   872

Regarding the limitations of our study: (i) The publicly available COVID-19 data   873
collections used are fairly small and may not encompass a wide range of disease pattern   874
variability. An appropriately annotated large-scale collection of CXRs with COVID-19   875

viral disease manifestations is necessary to build confidence in the models, improve their ₈₇₆ robustness, and generalization. (ii) The study is evaluated with the ROI annotations ₈₇₇ obtained from two expert radiologists. However, it would help to have more radiologists ₈₇₈ contribute independently in the annotation process and then arrive at a consensus that ₈₇₉ could reduce annotation errors. (iii) We used conventional convolutional kernels toward ₈₈₀ this study, however, future research could propose novel convolutional kernels that ₈₈₁ reduce feature dimensionality and redundancy and result in improved performance with ₈₈₂ reduced memory and computational requirements. (iv) Ensemble models require ₈₈₃ markedly high training time, memory, and computational resources for successful ₈₈₄ deployment and use. However, recent advancements in storage and computing solutions ₈₈₅ and cloud technology could lead to improvements in this regard. ₈₈₆

Conclusions ₈₈₇

In this study, we have demonstrated that a combination of repeated CXR-specific ₈₈₈ pretraining, fine-tuning, and ensemble learning helped in (a) transferring CXR-specific ₈₈₉ learned knowledge that is subsequently fine-tuned to improve COVID-19 detection in ₈₉₀ CXRs; and (b) improving classification generalization and localization performance by ₈₉₁ reducing prediction variance. Ensemble-based ROI localization helped in improving ₈₉₂ localization performance by compensating for the errors in individual constituent ₈₉₃ models. We also performed inter-reader variability analysis and program performance ₈₉₄ assessment by comparing them with a STAPLE-based estimated reference. This ₈₉₅ assessment highlighted the opportunity for improving performance through ensemble ₈₉₆ modifications, requisite parameter optimization, increased task-specific dataset size, and ₈₉₇ involving "truth" estimates from a larger number of expert collaborators. We believe ₈₉₈ that the results proposed are useful for developing robust models for tasks involving ₈₉₉ medical image classification and disease-specific ROI localization. ₉₀₀

Acknowledgment ₉₀₁

References

1. COVID-2019) situation reports. In: World Health Organization (WHO) Situation Reports. Coronavirus disease. 2020;.

2. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, et al. The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic. Chest. 2020;158(1):106–116.

3. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection; 2020. Available from: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection.

4. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran T. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology. 2020;.

5. Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK. Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays. IEEE Access. 2020;8:1–1.

6. Rajaraman S, Antani S. Weakly Labeled Data Augmentation for Deep Learning: A Study on COVID-19 Detection in Chest X-Rays. Diagnostics. 2020;10(6):358–358.

7. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology. 2020;.

8. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. Proceedings of the International Conference on Computer Vision (ICCV. 2017; p. 3462–3471.

9. Deng J, Dong W, Socher R, Li L, Li K, Li FF, et al. A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009; p. 248–255.

10. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering. 2017;19(1):221–248.

11. Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG. Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry. Medicine & Science in Sports & Exercise. 2017;49(9):1965–1973.

12. Zhao B, Tan Y, Bell DJ, Marley SE, Guo P, Mann H, et al. Exploring intra- and inter-reader variability in uni-dimensional, bi-dimensional, and volumetric measurements of solid tumors on CT scans reconstructed at different slice intervals. European Journal of Radiology. 2013;82(6):959–968.

13. Warfield SK, Zou KH, Wells WM. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. IEEE Transactions on Medical Imaging. 2004;23(7):903–921.

14. Diong J, Butler AA, Gandevia SC, Héroux ME. Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. PLOS ONE. 2018;13(8):e0202121.

15. Kim I, Rajaraman S, Antani S. Visual Interpretation of Convolutional Neural Network Predictions in Classifying Medical Image Modalities. Diagnostics. 2019;9(2):38–38.

16. Yadav O, Passi K, Jain CK. Using Deep Learning to Classify X-ray Images of Potential Tuberculosis Patients. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2018; p. 2368–2375.

17. Rajaraman S, Antani SK. Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. IEEE Access. 2020;8:27318–27326.

18. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology. 2017;284(2):574–582.

19. Rajaraman S, Sornapudi S, Kohli M, Antani S. Assessment of an ensemble of machine learning models toward abnormality detection in chest radiographs. Conf Proc IEEE Eng Med Biol Soc. 2019;2019:3689–3692.

20. Islam MT, Aowal MA, Minhaz AT, Islam KA; 2017.Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks. arXiv preprint arXiv: 170509850.

21. Zeiler MD, Fergus R; 2013.Visualizing and Understanding Convolutional Networks. arXiv preprint arXiv:13112901.

22. Dosovitskiy A, Brox T. Inverting Visual Representations with Convolutional Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; p. 4829–4837.

23. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; p. 2921–2929.

24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D, et al. Visual explanations from deep networks via gradient-based localization. Proceedings of the International Conference on Computer Vision (ICCV. 2017; p. 618–626.

25. Karim MR, Döhmen T, Rebholz-Schuhmann D, Decker S, Cochez M, Beyan O, et al; 2020. DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images. arXiv preprint arXiv:200404582.

26. Al-Khawari H, Athyal RP, Al-Saeed O, Sada PN, Al-Muthairi S, Al-Awadhi A. Inter- and intraobserver variation between radiologists in the detection of abnormal parenchymal lung changes on high-resolution computed tomography. Annals of Saudi Medicine. 2010;30(2):129–133.

27. Jiang Y, Guo D, Li C, Chen T, Li R. High-resolution CT features of the COVID-19 infection in Nanchong City: Initial and follow-up changes among different clinical types. Radiology of Infectious Diseases. 2020;.

28. Balabanova Y, Coker R, Fedorin I, Zakharova S, Plavinskij S, Krukov N, et al. Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study. BMJ. 2005;331(7513):379–382.

29. Kermany DS, Goldbaum M, Cai W, Valentim C, Liang H, Baxter SL. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122–1131.

30. Shih G, Wu CC, Halabi SS, Kohli MD, Prevedello LM, Cook TS. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. Radiol Artif Intell. 2019;1(1):180041–180041.

31. Irvin J, Rajpurkar P, Ko M, Yu Y, Silviana CI, Chute C. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the 33rd AAAI conference on artificial intelligence (AAAI). 2019; p. 590–597.

32. Cohen JP, Morrison P, Dao L;. 2020. COVID-19 image data collection. arXiv preprint arXiv:200311597.

33. Hesamian MH, Jia W, He X, Kennedy P. Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. Journal of Digital Imaging. 2019;32(4):582–596.

34. Baldi P, Sadowski P. The dropout learning algorithm. Artificial Intelligence. 2014;210:78–122.

35. Candemir S, Antani S. A review on lung boundary detection in chest X-rays. International Journal of Computer Assisted Radiology and Surgery. 2019;14(4):563–576.

36. Dutta A, Zisserman A. The VIA Annotation Software for Images, Audio and Video. Proceedings of the 27th ACM International Conference on Multimedia (MM). 2019; p. 2276–2279.

37. Zerhouni E, Lanyi D, Viana MP, Gabrani M. Wide residual networks for mitosis detection. Proceedings of the IEEE International Symposium on Biomedical Imaging. 2017; p. 924–928.

38. Zhang HX, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the International Conference on Computer Vision (ICCV). 2016; p. 770–778.

39. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the International Conference on Learning Representations (ICLR). 2015; p. 1–14.

40. Chollet F. Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; p. 1251–1258.

41. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; p. 2818–2826.

42. Liu HZ, Maaten LVD, Weinberger KQ. Densely connected convolutional networks. Proceedings of the International Conference on Computer Vision (ICCV. 2017; p. 4700–4708.

43. Sandler M, Howard AG, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; p. 4510–4520.

44. Pham MY, Zoph GB, Le QV, Dean J. Efficient neural architecture search via parameter sharing. Proceedings of the International Conference on Machine Learning (ICML). 2018; p. 4092–4101.

45. Zahery M, Maes HH, Neale MC. CSOLNP: Numerical Optimization Engine for Solving Non-linearly Constrained Problems. Twin Research and Human Genetics. 2017;20(4):290–297.

46. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. Proceedings of the European Conference on Computer Vision (ECCV). 2014; p. 740–755.

47. Kao LS, Green CE. Analysis of Variance: Is There a Difference in Means and What Does It Mean? Journal of Surgical Research. 2008;144(1):158–170.

48. Platzer A, Polzin J, Rembart K, Han PP, Rauer D, Nussbaumer T. BioSankey: Visualization of Microbial Communities Over Time. Journal of Integrative Bioinformatics. 2018;15(4).

49. Acuff NV, Linden J. Using Visualization of t-Distributed Stochastic Neighbor Embedding To Identify Immune Cell Subsets in Mouse Tumors. The Journal of Immunology. 2017;198(11):4539–4546.

List of Figures

Fig 1. The architecture of the custom U-Net with dropout.

Fig 2. Segmentation workflow showing UNet-based mask generation and lung ROI cropping.

Fig 3. The workflow of the proposed repeated CXR-specific pretraining and fine-tuning.

X_i

Conv 5×5

Dropout (0.3)

Conv 5×5

X_{i+1}

Fig 4. A custom Wide Residual Network (WRN) with dropout regularization.

Fig 5. The architecture of the CNNs used in the first stage of CXR-specific pretraining. (I/P = Input, I-CNN = truncated ImageNet-pretrained CNNs, ZP = Zero-padding, CONV = Extra convolution layer, GAP = Global Average Pooling, DO = Dropout, D = Final dense layer with Softmax activation)
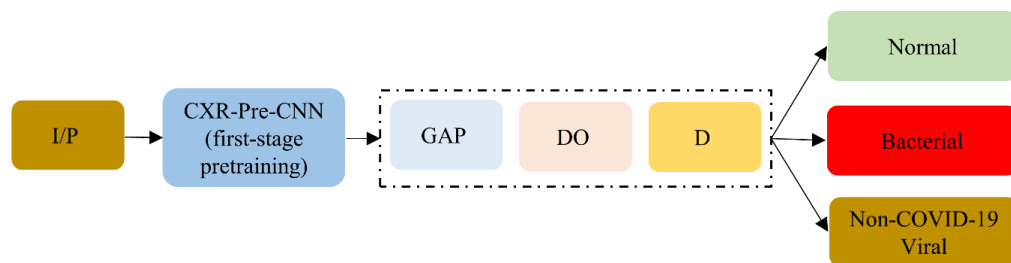
Fig 6. The architecture of the CNNs used in the second stage of pretraining. (I/P = Input, CXR-Pre-CNN = CXR-specific CNNs from the first stage of pretraining, truncated at their deepest convolutional layer, GAP = Global Average Pooling, DO = Dropout, D = Final dense layer with Softmax activation.
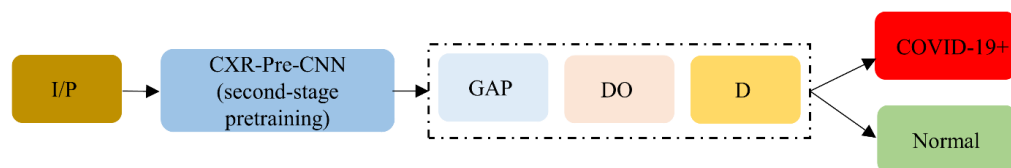
Fig 7. The architecture of the CNNs fine-tuned toward COVID-19 detection. (I/P = Input, CXR-Pre-CNN = CXR-pretrained CNNs from the second stage of pretraining, truncated at their deepest convolutional layer, GAP = Global Average Pooling, DO = Dropout, D = Final dense layer with Softmax activation.

(A)   (B)   (C)   (D)
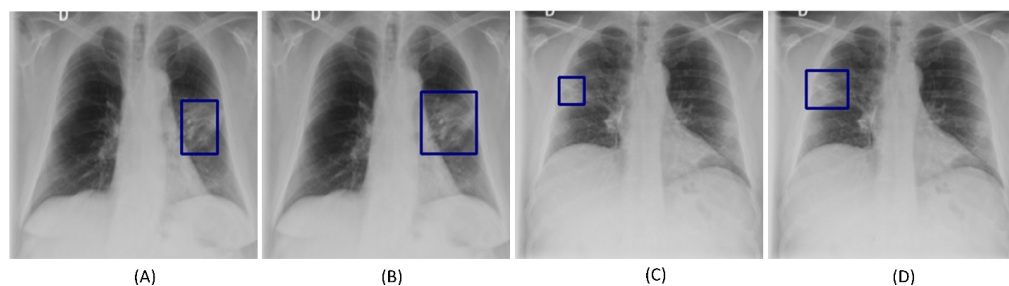
Fig 8. Examples showing inter-reader variability in annotating COVID-19 disease ROI. (A) and (B) show the annotations (bounding boxes in blue) of Rad-1 and Rad-2, respectively, for a given COVID-19 disease labeled image; (C) and (D) shows the GT annotations of Rad-1 and Rad-2, respectively for another COVID-19 disease labeled image.
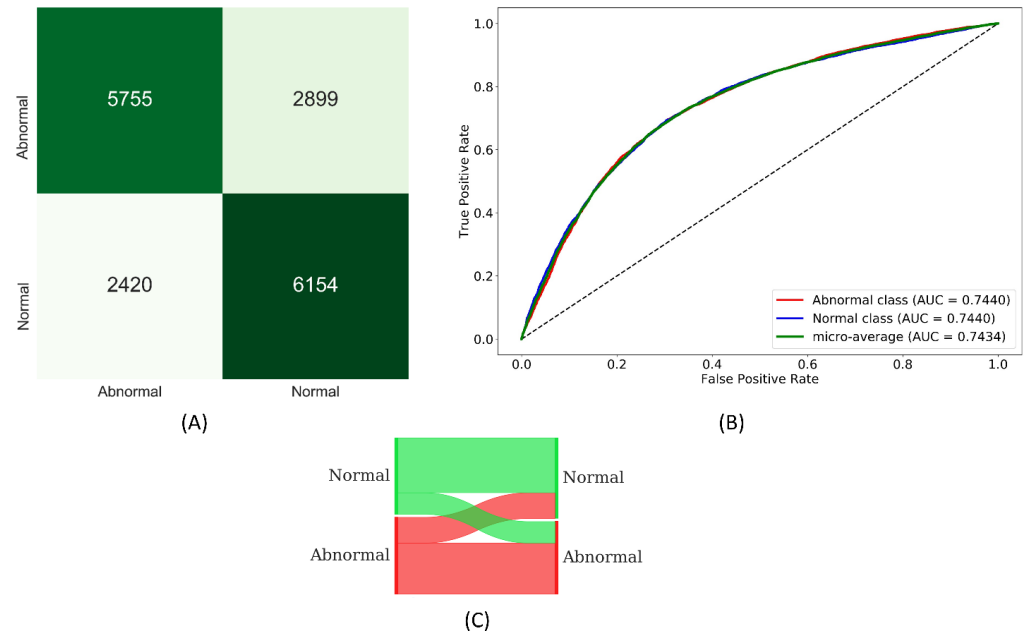
Fig 9. (A) Confusion matrix; (B) ROC curves; (C) Normalized Sankey flow diagram obtained using the VGG-19 model during the first-stage of CXR-specific pretraining.
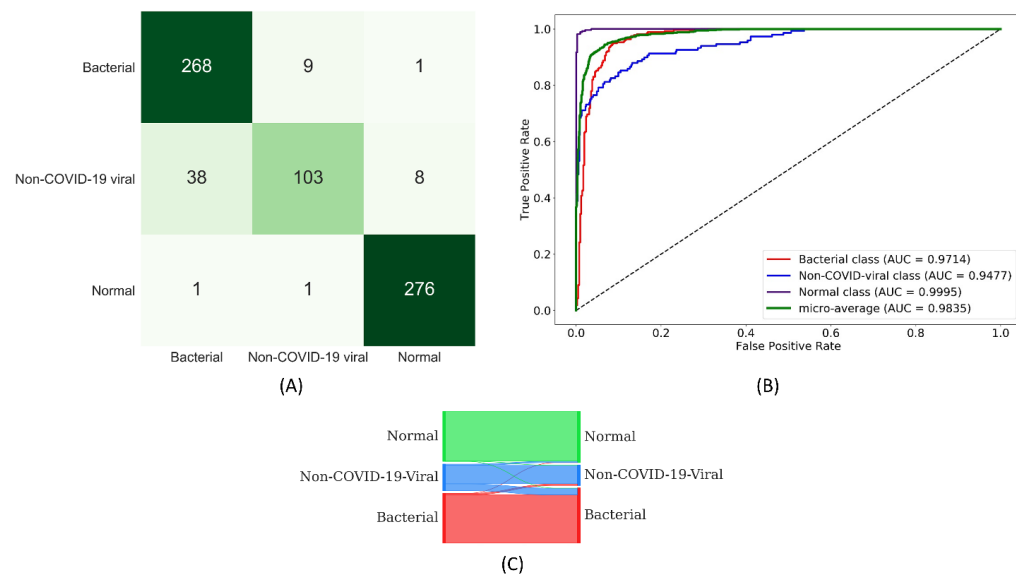
Fig 10. **(A) Confusion matrix; (B) ROC curves; (C) Normalized Sankey flow diagram obtained using the DenseNet-121 model during the second stage of CXR-specific pretraining.**
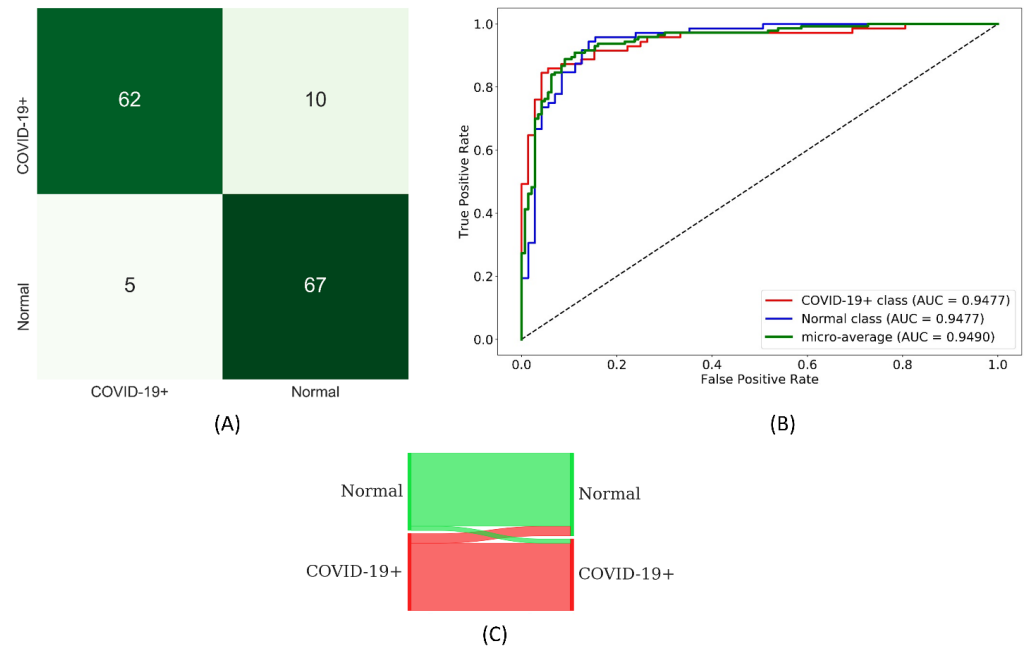
Fig 11. (A) Confusion matrix; (B) ROC curves; (C) Normalized Sankey flow diagram obtained using the ResNet-18 model during fine-tuning for COVID-19 detection.
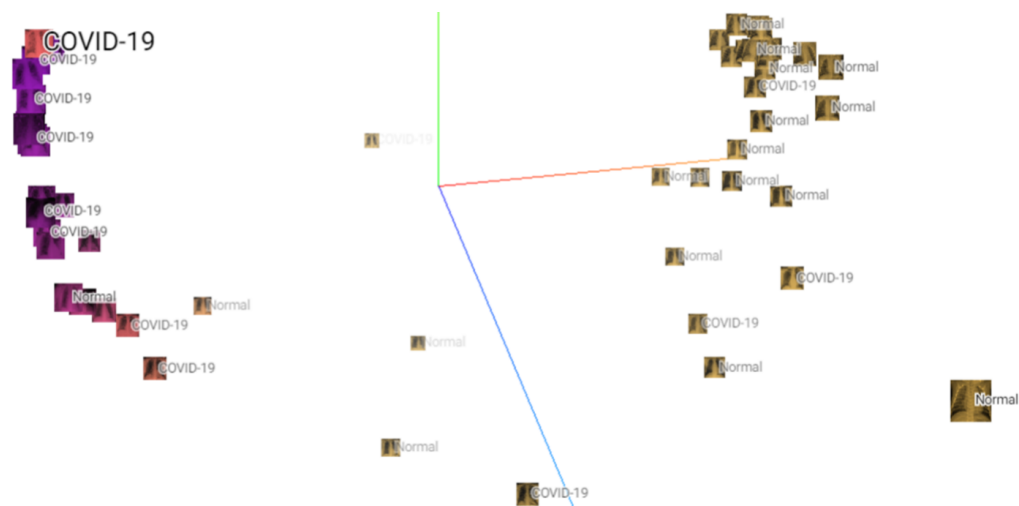
Fig 12. Visualizing feature embedding for the ResNet-18 fine-tuned model using t-SNE. The plot shows a 1024-dimensional space embedded into 2 dimensions.
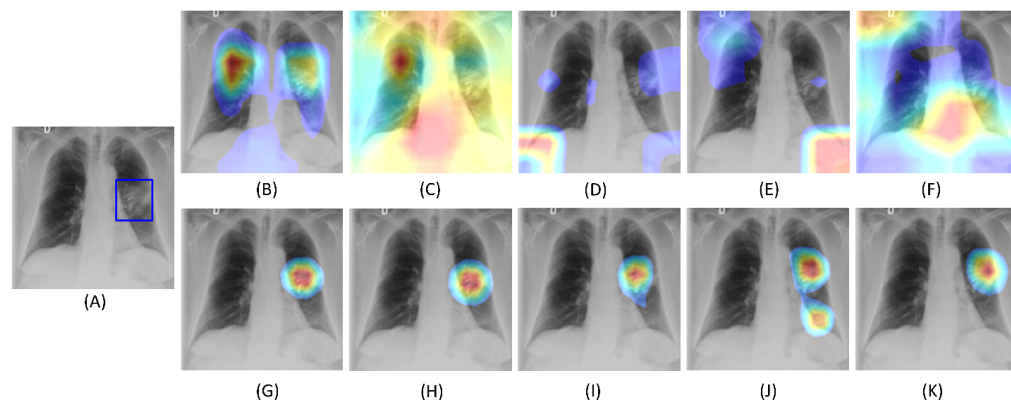
Fig 13. COVID-19 viral disease ROI CRM-based localization achieved using the fine-tuned models and their baseline counterparts. (A) Original CXR with STAPLE-generated consensus ROI (shown in blue); (B) Baseline VGG-16; (C) Baseline VGG-19; (D) Baseline MobileNet-V2; (E) Baseline DenseNet-121; (F) Baseline Inception-V3; (G) Fine-tuned VGG-16; (H) Fine-tuned VGG-19; (I) Fine-tuned MobileNet-V2; (J) Fine-tuned DenseNet-121; (K) Fine-tuned Inception-V3.
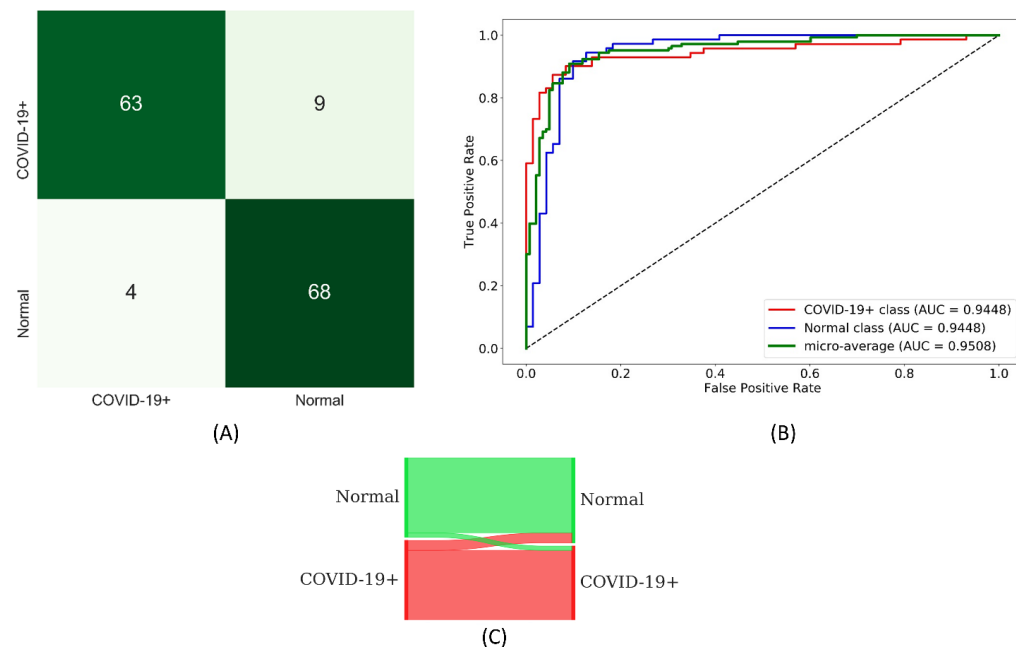
Fig 14. **(A) Confusion matrix; (B) ROC curves; (C) Normalized Sankey flow diagram obtained through weighted averaging of the top-3 fine-tuned CNNs toward COVID-19 detection.**
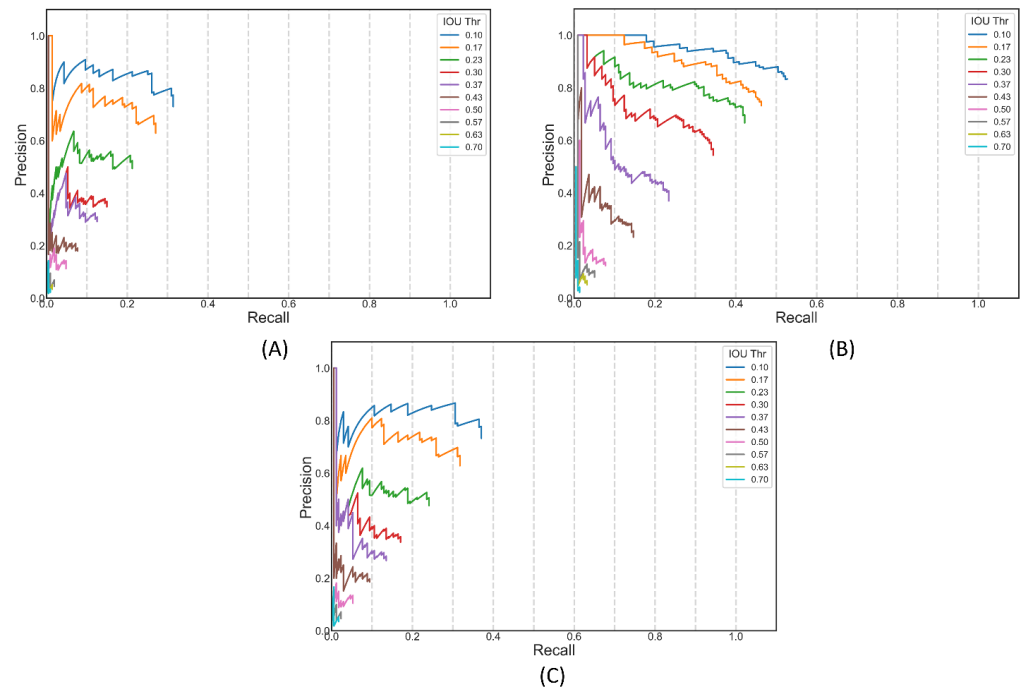
Fig 15. PR curves for the top-performing models using annotations of (A) Rad-1; (B) Rad-2; (C) STAPLE-generated ROI consensus.
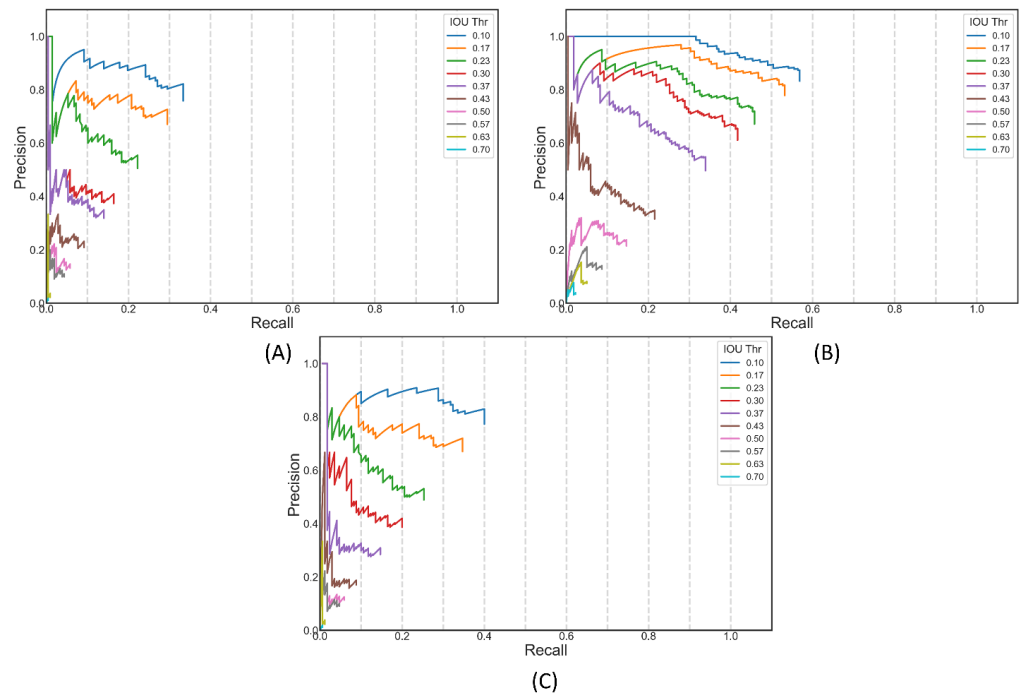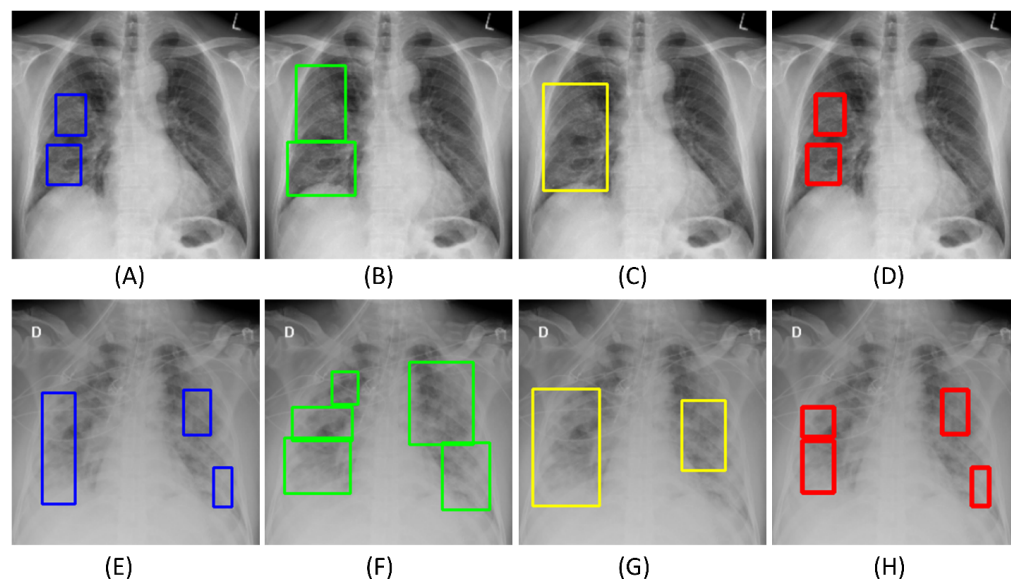
Fig 16. PR curves for the top-N performing models using annotations of (A) Rad-1 (N = 3); (B) Rad-2 (N = 5); (C) STAPLE-generated consensus ROI (N = 3).

Fig 17. Sample CXRs showing annotations generated by Rad-1 (A) and (E) (in blue); Rad-2 (B) and (F) (in green); (C) and (G) top-3 ensemble using STAPLE-generated consensus ROI (program) (in yellow); (D) and (H) STAPLE-generated consensus ROI annotation (in red).
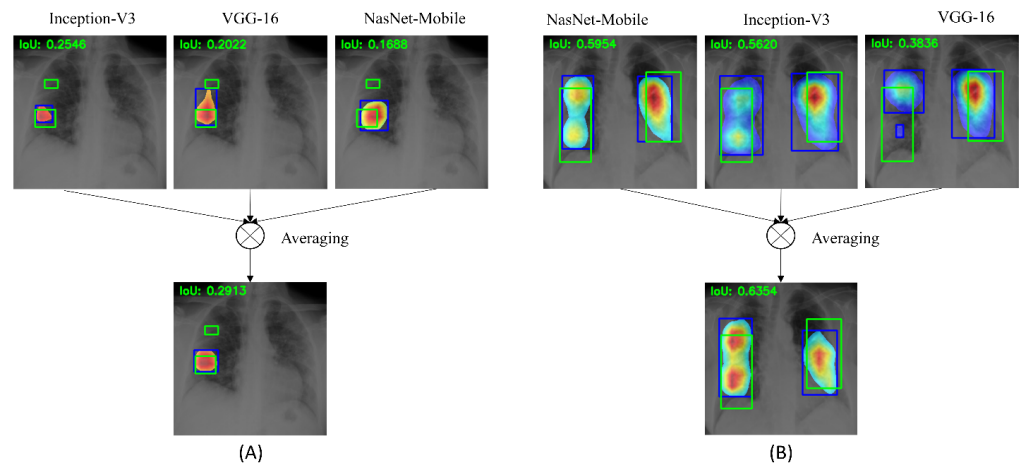
Fig 18. Instances of ensemble CRMs combining (A) top-3 CNNs using STAPLE-generated consensus ROI annotation; (B) top-5 CNNs using Rad-2 annotations. The green box denotes reference ROI annotation and the blue box denotes ensemble CRM localization.
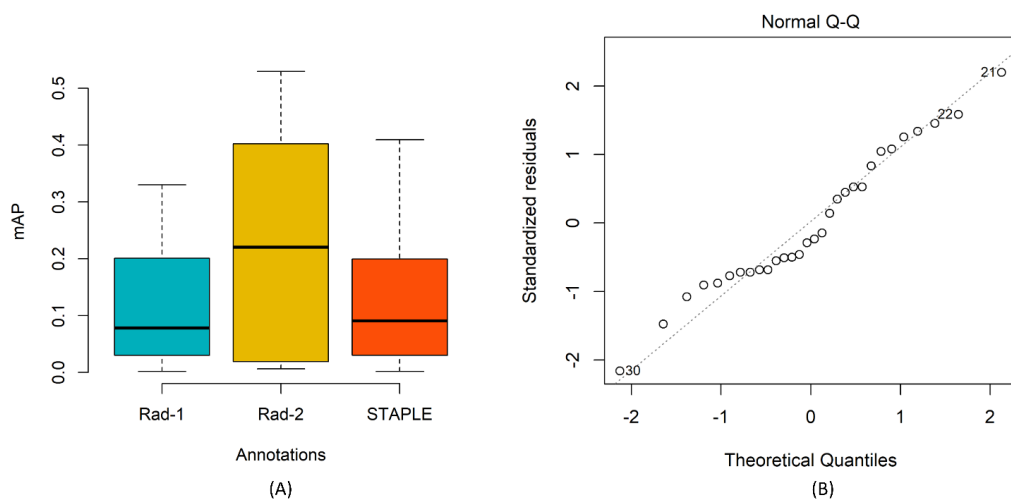
Fig 19. (A) Mean plot for the mAP scores obtained by the top-N ensembles using Rad-1, Rad-2, and STAPLE-generated consensus ROI annotations; Error bars represent standard errors. The differences are not statistically significant; (B) Residual plot showing the data follow the normal distribution .
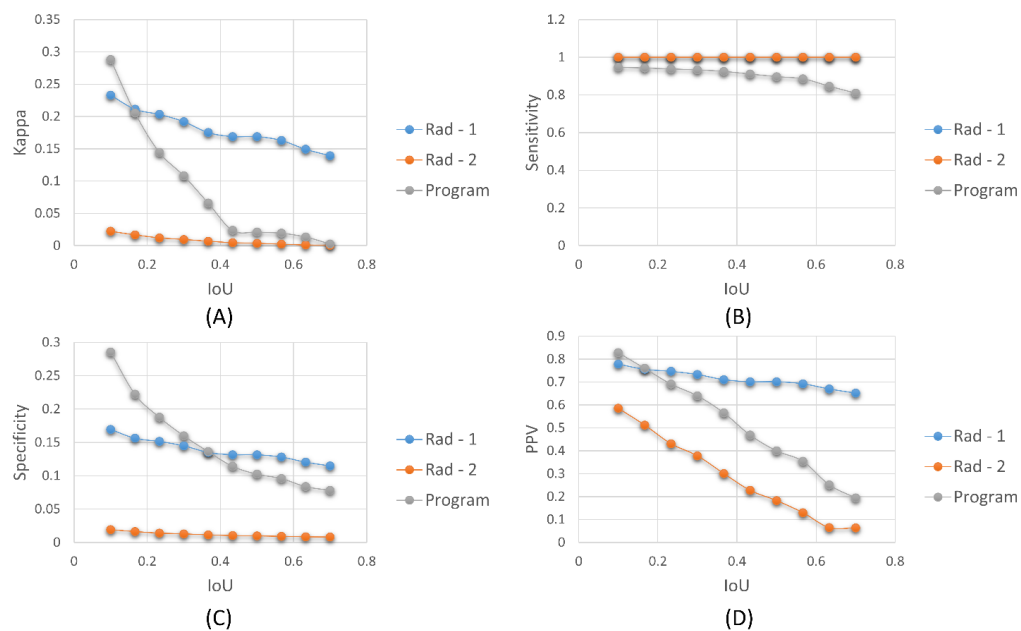
Fig 20. Assessing inter-reader variability and program performance. The
following performance metrics are measured and plotted for 10 different IoU thresholds
in the range $(0.1 - 0.7)$: (A) Kappa statistic; (B) Sensitivity; (C) Specificity; (D) PPV.