



Multi-head deep learning framework for pulmonary disease detection and severity scoring with modified progressive learning

Asad Mansoor Khan ^a, Muhammad Usman Akram ^{a,*}, Sajid Nazir ^b, Taimur Hassan ^c,
Sajid Gul Khawaja ^a, Tatheer Fatima ^d

^a National University of Sciences and Technology, Islamabad, 44000, Pakistan

^b Department of Computing, Glasgow Caledonian University, Glasgow, UK

^c Center for Cyber-Physical Systems (C2PS), Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi 127788, United Arab Emirates

^d Resident Radiologist, Pakistan Institute of Medical Sciences, Islamabad, Pakistan

ARTICLE INFO

Keywords:

Medical imaging

Severity grading

Segmentation

Classification

Multi-head convolutional neural network

ABSTRACT

Chest X-rays (CXR) are the most commonly used imaging methodology in radiology to diagnose pulmonary diseases with close to 2 billion CXRs taken every year. The recent upsurge of COVID-19 and its variants accompanied by pneumonia and tuberculosis can be fatal in some cases and lives could be saved through early detection and appropriate intervention for the advanced cases. Thus CXRs can be used for an automated severity grading of pulmonary diseases that can aid radiologists in making better and informed diagnoses. In this article, we propose a single framework for disease classification and severity scoring produced by segmenting the lungs into six regions. We present a modified progressive learning technique in which the amount of augmentations at each step is capped. Our base network in the framework is first trained using modified progressive learning and can then be tweaked for new data sets. Furthermore, the segmentation task makes use of an attention map generated within and by the network itself. This attention mechanism allows to achieve segmentation results that are on par with networks having an order of magnitude or more parameters. We also propose severity score grading for 4 thoracic diseases that can provide a single-digit score corresponding to the spread of opacity in different lung segments with the help of radiologists. The proposed framework is evaluated using the BRAX data set for segmentation and classification into six classes with severity grading for a subset of the classes. On the BRAX validation data set, we achieve F1 scores of 0.924 and 0.939 without and with fine-tuning, respectively. A mean matching score of 80.8% is obtained for severity score grading while an average area under receiver operating characteristic curve of 0.88 is achieved for classification.

1. Introduction

Radiologists rely on Chest X-rays (CXR) as they are an easily accessible, low-cost method that can be brought to the patient in certain situations and therefore can prove vital for pulmonary disease diagnosis [1,2]. COVID-19 pandemic has overwhelmed radiologists because of an influx in the number of patients as, now, they have to deal with the unprecedented challenges of diagnosing a significantly large number of CXRs [3]. While it is true that COVID-19 pandemic has put significant strain on an already stretched medical system, even before the pandemic, the CXR evaluation and interpretation demands far exceeded the number that could be handled by the radiologists [4,5]. Although, Computed Tomography (CT) scans are more sensitive than

CXR and better for diagnosis [6] and have been used for classification of COVID-19 as well [7,8], the ubiquity of the CXR makes it far more practical as a diagnosis tool along with the added benefit of less exposure to radiation. This is one of the reasons that several, large, public CXR data sets have been made available for both classification [9–11] as well as segmentation [12,13]. Similarly, recently, data sets for severity scoring [14,15] have been made available as well. However, the limitation of these severity-scoring data sets stems from the fact that they provide severity scoring just for COVID-19 cases. The utility of publicly available, extensively labelled CXR data sets can sometimes be hampered due to the need of specialised annotations by the radiologists [16].

* Corresponding author.

E-mail addresses: asad.mansoor@ceme.nust.edu.pk (A.M. Khan), usman.akram@ceme.nust.edu.pk (M.U. Akram), sajid.nazir@gcu.ac.uk (S. Nazir), taimur.hassan@ku.ac.ae (T. Hassan), sajid.gul@ceme.nust.edu.pk (S.G. Khawaja).

<https://doi.org/10.1016/j.bspc.2023.104855>

Received 14 October 2022; Received in revised form 4 March 2023; Accepted 11 March 2023

Available online 24 March 2023

1746-8094/© 2023 Elsevier Ltd. All rights reserved.

Similar to its use in other domains, Convolutional Neural Networks (CNN) have been used for automated classification and segmentation for CXRs and have become state-of-the-art methods for the aforementioned tasks. As a result, many researchers have employed a variety of CNN architectures for the classification of different pathologies [17–19], for segmentation of lungs from the CXR [20–22] and severity scoring for COVID-19 progression [14,15]. The more generalisable such systems are, the more widespread they can be.

[23] evaluated the performance of six different neural network architectures used for binary classification of CXR images for pneumonia. The researchers documented the effect of transfer learning using CNN architectures of various depths and complexity for a binary classification problem. Their work showed that the number of trainable parameters is not directly related to the performance of the network. [24] initially trained the model using a patch-based strategy that focused on the anomaly and then used transfer learning to train the same architecture on entire X-ray images. The initial patch-based learning was carried out using various amounts of data, with the use of 100% data producing the best results. Gradient Class Activation maps were used to localise abnormalities as well. In order to make use of the free-text reports available with the CXR, [25] leveraged a Recurrent Neural Network (RNN) in combination with a CNN in TieNet by using a combination of text embeddings and image features. Using the text reports as a priori knowledge to generate attention maps, the authors combined multi-level attention models into a single framework. This not only improved the baseline Area Under the Curve (AUC) scores but also allowed the researchers to re-purpose TieNet purely for report generation as well. [26] introduced an attention mechanism by combining two CNN architectures; one of which examined the CXR globally while the other inspected a more localised region. The local region of interest was identified by global CNN. Both the networks were trained in tandem and their feature space was combined in order to train them using a shared loss. The researchers validated their methodology on the ChexX-ray 14 data set [25]. Vision Transformer (ViT) [27] was more recently repurposed by [15] for the classification of COVID-19. To enable the framework to include the low-level corpus findings, the researchers first trained the transformer architecture using a sizable, publicly accessible data set. The researchers also carried a COVID-19 severity grading using the same transformer as a backbone. Performance significantly improved because of the combination of vision transformer and the large-scale data set.

[28] employed a patch-based classifier relying on segmentation masks from a DenseNet103 CNN in order to utilise the few images available for COVID-19 detection. The researchers observed that the network tended to yield under-segmentation when applied to abnormal CXRs due to the considerable disparity between the available normal and abnormal images. [29] used a pair of CNNs with the initial network doing patch-wise classification into the lung and no lung regions to enhance segmentation performance on anomalous images. The reconstruction network generated a mask from these binary categorised patches, creating a new mask. The final mask was the combination of the two masks generated. A few-shot semantic segmentation approach with sparse labelled images was proposed and evaluated with four CXR data sets for grid and points annotations. The proposed methods used dense labels in meta-test and used sparse labels in the meta-learning to predict dense labels from sparse ones [16]. Chronic Obstructive Pulmonary Disease (COPD) is usually accompanied by lung lesions, the segmentation of which can help diagnose the problem and its extent. For the segmentation of lesions in CT scans, [30] used a novel Renyi's entropy multi-threshold image segmentation based on an enhanced Slime Mould Algorithm.

While there are several thoracic disorders that can vary on a severity spectrum and associated severity scoring systems [31–33], recent severity scoring methodologies have only focused on COVID-19 [14,15,34]. While the approaches for other disorders may rely on physical examination or questionnaire [32,33], the severity grading

schemes for COVID-19 only require CXR. The ability to predict the severity score from a single CXR provides for a fast evaluation of the level of care needed for the patient and can have a big impact on how the patient is treated.

Even though, many techniques and frameworks have been proposed for the classification of different lung diseases and opacity localisation, there is a lack of a single framework that not only classifies a CXR image in a particular disease class but also segments the lungs' opacity regions. Furthermore, while COVID-19 has been the focus of recent severity classification attempts, a lack of severity classification for different pathologies still exists. To address this gap, our main research contributions in this article are as follows:

1. We propose a single framework consisting of a disease classification and severity grading for different diseases.
2. A segmentation network is utilised that despite its relatively small size is able to perform relatively close to large architectures such as U-Net.
3. We experimentally show that while good performance can be achieved in segmentation using publicly available data sets, fine tuning on just a small number of samples from the target data set can actually improve the segmentation performance even further.
4. We provide segmentation masks with severity grades for a validation data set that has been validated by a human radiologist.

The remaining paper is organised as follows: related work is described in Section 2. Section 3 Materials and Methods describes the data set and the proposed architecture. The results are given in Section 4. Section 5 provides the discussion of the results and finally Section 6 concludes the paper.

2. Related work

Our proposed framework is primarily focused on three components: the first component uses progressive learning for image-level classification, the second component segments the lungs, and the third and final component determines severity scores for various pulmonary diseases based on lung opacification. With this in mind, the current section offers an overview of the relevant literature.

Using neural architecture search and scaling [35] created the EfficientNet v2 family of CNN architectures, which improved parameter efficiency and training speed. In addition, the researchers proposed a pyramid-style strategy to train models with progressively larger input sizes using progressive learning with adaptively modifying regularisation. They showed that increasing regularisation with each input size step can solve the problem of performance degradation. The proposed model beat earlier models and was also faster and more efficient. The family of EfficientNetv2 classifiers has been leveraged for classification of COVID-19 owing to their fewer parameters and comparatively smaller size. For COVID-19 diagnosis using CXR data, a classification network namely DFFCNet was proposed. For feature extraction, the model used the backbone network EfficientNetV2. In comparison to the other chosen models, the proposed framework performed better in experiments [33].

COVID-19 was detected using pre-trained models like Xception, InceptionV3, and EfficientNetV2 from CXR and CT images. EfficientNetV2 with fine tuning produced the best performance for the CXR data set, whereas the LightEfficientNetV2 model produced the highest performance for the CT data set [36]. Several pre-trained models including EfficientNet B3, EfficientNet V2, HrNet and ResNet50d were used for an automatic diagnosis of Myocarditis in Cardiac Magnetic Resonance (CMR) images. The EfficientNetV2 had the best performance compared to the other pre-trained models [37].

Most of the segmentation methodologies rely on a U-Net like architecture with an Encoder–Decoder configuration. Little W-Net model was proposed for artery/vein segmentation problem with minimalistic

models consisting of two conventional U-Nets [38], based on the notion in [39] using the aggregate loss of each U-Net in training. The suggested model was tested on 10 different data sets and shown to be superior to earlier techniques, requiring 1–3 orders of magnitude less computing power than previous CNNs [38]. The researchers were able to demonstrate that the key benefit of such a design is the network's capacity to generate an attention map that can aid in improving the final segmentation mask. For many medical applications, this type of network architecture has proven to have good performance. The narrow vascular anatomy and atrophic modifications make binary vessel segmentation difficult. [40]. AutoMorph was suggested for the fundus images as a method of automating retinal morphology analysis. In order to minimise the large segmentation errors, the authors employed the Little W-Net architecture [40]. The reliability and accuracy of the size is crucial for polyp size measurements in colonoscopy images. The segmentation was performed with Little W-Net architecture due to its reduced number of parameters which was accompanied by increased feature representation [41]. To the best of our knowledge, however, this approach has never been applied for lung segmentation.

Researchers have devised many scoring systems to provide a quantified score for the severity of certain disorders. Most methods has been to divide the lung into discrete areas and assign a score to each zone as well as an overall score. [42] assigned a grade to each lung individually, but [43] separated the lungs into four quadrants. [14] refined the method by separating each lung into three areas for a total of six.

Severity assessment can help medical care providers come up with a better treatment plan. Keeping this in view, the severity assessment of COVID-19 CXR images into mild, moderate, severe, and critical with CNN was proposed by [44]. The study utilised nine publicly available CXR data sets with 3260 images in total. The disease severity score was based on an opacity score by two radiologists.

[14] devised the Brixia score, which awarded a score to each of the six lung regions acquired by splitting the lungs into three equal parts on a scale of 0 to 3. The level of lung compromise in COVID-19 cases was reflected by a score for each location, with 0 indicating no abnormalities and 3 indicating significant aberrations. They also proposed BS-Net, which used a pyramid technique to integrate features collected at multiple scales after employing a multi-feature area aligner to automatically align the segmentation masks using a multi-feature region aligner and then used a series of convolutional blocks to map the input feature to the final Brixia Score.

[15] proposed a vision transformer-based framework for COVID-19 diagnosis and severity quantification in a multi-task learning approach by employing a large CXR data set for training the backbone model so that it can learn low-level generalised features. The deep features from the backbone model were then used in conjunction with the vision transformer for the prediction of disease class and severity map for severity quantification. The researchers used a similar scoring technique as [14], however, they modified the scoring system such that each of the six lung regions could only have a maximum score of 1 with a total score of 6 for the entire CXR. The use of vision transformer achieved comparable performance to the state-of-the-art with a unified framework.

3. Materials and method

This section describes the data set, image pre-processing and augmentation techniques, proposed CNN model and the application architecture.

3.1. Data sets

As the proposed framework performs classification, segmentation and severity grading, so the use of data sets can also be divided to these three categories.

Table 1

Distribution of different classes in the data set that have been used for classification.

Classes	CheXpert [9]	BRAX [10]
Cardiomegaly	25673	2171
Consolidation	14832	1678
Edema	49675	25
No Finding	16974	2000
Pleural Effusion	76899	1006

Table 2

Distribution of samples with masks in the data sets.

Data set	Training data	Validation data
JSRT [12]	198	49
Montgomery [13]	111	27
Shenzhen [13]	453	113

3.1.1. Data sets for classification

BRAX consists of 24,959 CXR exams and 40,967 images from a Brazilian hospital [10]. The images were extracted from Picture Archiving and Communication System (PACS) from the Hospital Israelita Albert Einstein (HIAE). Similar to [9], the labels in BRAX for 14 conditions can be positive, negative, or uncertain. Out of the 14 pathologies, we selected anterior–posterior (AP) or posterior–anterior (PA) images of 6 pathologies namely atelectasis (AT), cardiomegaly (CA), consolidation (CO), edema (ED), no finding (NF) and pleural effusion (PE) that have been marked as positive. In BRAX, images belonging to No Finding far outnumber the samples of all other classes. Therefore, in the interest of maintaining a class balance when using the BRAX dataset, downsampling was applied to this class by randomly selecting a subset of the images. Fig. 1 illustrates the presentation of a selection of diseases from the BRAX [10] dataset. It also highlights the possibility that due to the similarity of symptoms, several lung diseases may be hard to distinguish from one another.

It is also worth noting that there is no validation data set for BRAX that has been provided by the authors. Table 1 summarises the classification data set resources used for this research work.

3.1.2. Data sets for segmentation

Table 2 summarises data set resources used for training the segmentation model. Public data sets [12,13] were used for training and validation of the framework. The data from all the data sets were combined and were treated as a single data set.

As the BRAX data set does not contain a validation subset of any kind, therefore we created a validation subset for segmentation consisting of 200 images from BRAX that were manually annotated by a senior radiologist.

3.1.3. Data sets for opacity localisation

SIIM-FISABIO-RSNA COVID-19 detection data set was made accessible through a Kaggle [45] public challenge. The goal of this data set is to locate lung opacity regions in the CXR images when COVID-19 and associated pneumonia types are detected. There are 6336 total images in the training data set, with resolutions ranging from 846×1353 to 4891×4020 . Only 4224 of the total number of images have opacity annotations. The opacity localisation head was trained using this subset.

3.1.4. Data curation for severity quantification

Opacification is caused by the displacement of air from the lungs as a result of different causes which can show up in CXR images [46]. While for COVID-19, opacification of the lungs is a clear indication of presence and progression, opacities also occur in different pulmonary diseases [47–51]. Keeping this in view, we have devised a severity scoring system for four lung pathologies: Atelectasis, Consolidation,

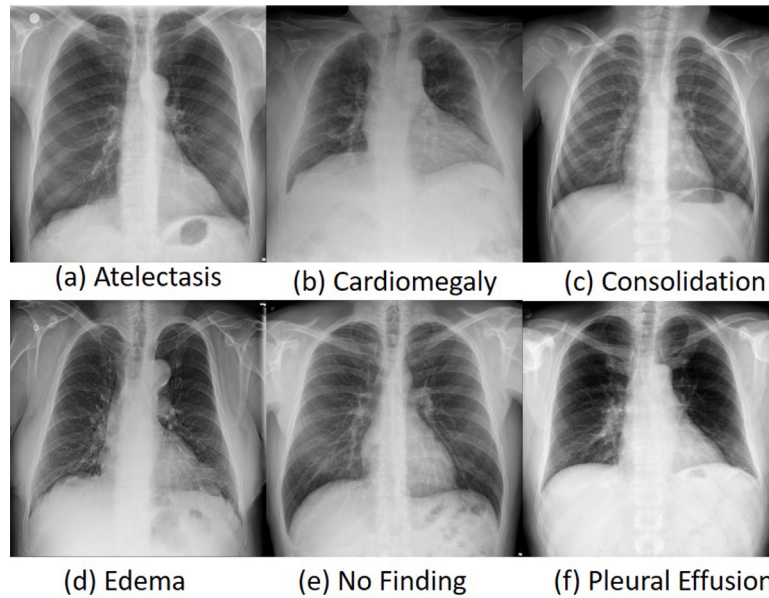


Fig. 1. Example images from (a) to (e) for the various pulmonary diseases in the data set.

Edema, and Pleural Effusion. Adapting the methodology of the division of both lungs into three subsections for a total of six regions from [14], we have simplified the opacity scoring to just binary as done by [52]. In essence, each of the six regions can have a score of 0 or 1 depending on the absence or presence of opacity. The scores for all six regions are added up to achieve a final severity score for both lungs.

A validation subset of 100 images from BRAX [10] has been manually annotated for the presence of opacities in the six lung regions by the same radiologist keeping in view the aforementioned methodology of the scoring system. We have validated the performance of our framework on this subset.

3.2. Proposed framework

Our proposed multi-output framework allows for multiple insights from a single CXR image by the means of different heads as detailed in Fig. 2. The classification head through an ensemble of multiple CNN backbones outputs an image-level classification probability of different pulmonary pathologies. The lung segmentation mask obtained from the segmentation head is used in combination with opacity localisation to define the severity of a subset of pathologies that our framework has been trained for. The sections below describe the different heads in detail.

3.2.1. Image segmentation

Isolating the lungs in a CXR is necessary whenever there is a need to provide a conclusion that relies on the specific sub-regions of the lungs. In the W-Net architecture [38], two UNets are strung together to form the W shape in which the output of the first Unet is concatenated with the original image before passing it to the second one. The UNet, generally, can be parameterised by two values: depth d and number of filters f in the first convolutional layer and can be represented as $\phi_{f,d}$ and the output from such an architecture is denoted as $y = \phi_{f,d}(x)$. Here $\phi_{f,d}^n$ denotes the architecture of a UNet where $n = [0, 1]$, y denotes the output or the *feature map* from such an architecture and $Loss(\phi_{f,d}^n)$ denotes the Categorical Crossentropy loss for a particular model n . For every succeeding convolutional layer, the number of filters in that layer is doubled until the defined depth is reached. The final output from a WNet can be represented as Eq. (1).

$$y = \phi_{f,d}^2(x, \phi_{f,d}^1(x)) \quad (1)$$

The number of parameters in this type of architecture are a function of the depth and filters in the first convolutional layers which is signified by the “Little” moniker in the Little WNet. Increasing both these parameters results in a greater number of trainable parameters of the architecture. Fig. 3 shows the detailed architecture of the segmentation head being used.

During training, the loss minimisation is performed on the linear summation of the losses of both the UNets ($Loss = \frac{1}{2}(Loss(\phi^1)) + \frac{1}{2}(Loss(\phi^2))$) which is then used for back propagation using the Adam optimiser. The training strategy makes use of a decaying cyclic cosine annealing scheduler which is a combination of a cosine annealing scheduler and a decaying learning rate scheduler. It varies the learning rate between 10^{-3} and 10^{-6} over a period of 10 epochs. The data set used for training is a combination of all segmentation [12,13] data sets. This ensures that the model is machine agnostic i.e. the model performs equally well for any CXR irrespective of the capturing machine.

Images are resized to 512×512 resolution and a single channel is kept at the input resulting in a 2-channel image for the second UNet where the attention map output from the initial Unet constitutes the 2nd channel. Augmentations include just the horizontal flip and cut-out [53] while the batch size is two. In order to gauge the performance of the model during the training, the F1 score is used that is computed for the validation split and the best-performing model is kept.

3.2.2. Image classification

Image level classification tends to be an important aspect of any framework as it provides an overall diagnosis for the image. In order to incorporate the progressive learning [35] strategy in the proposed framework, the classification head consists of 4 backbones based on EfficientNet B0 [54] at varying sizes that go from 256 all the way up to 768. Progressive learning proposes that weights from the previous, smaller network are used as initial weights for the next, larger network instead of random initialisation with increasingly aggressive regularisation at each stage as a means of making the examples harder for the network. However, in contrast to the original idea of increasingly aggressive regularisation, in our training strategy, the augmentation for each progressive stage have been capped at the same level as that of the first stage. Fig. 4 highlights the difference between the progressive learning strategy proposed by [35] and the changes made to it. Fig. 5 shows the architecture of the classification head of the framework.

As mentioned earlier, six classes namely Atelectasis, Cardiomegaly, Consolidation, Edema, No Finding and Pleural Effusion were selected

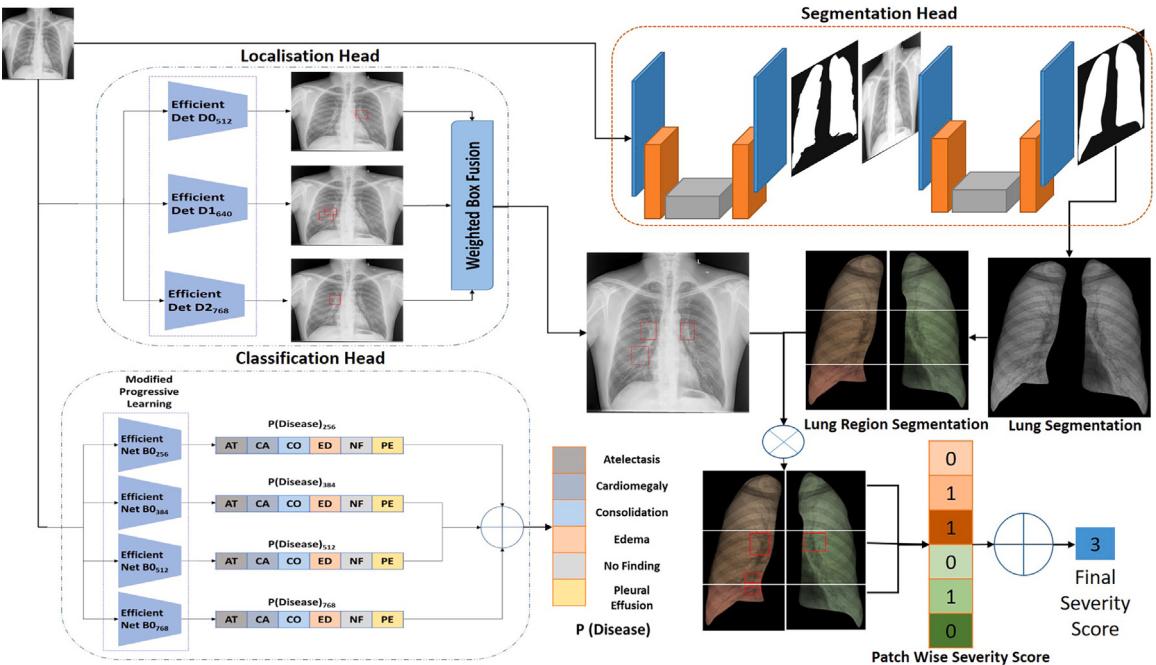


Fig. 2. Proposed System Architecture. The classification head (Bottom Left) that outputs the probability of a disease from an ensemble. Opacities are first localised using the localisation head (Top Left) and are then combined with the different lung regions obtained using the segmentation head to obtain the final severity score (Bottom Right).

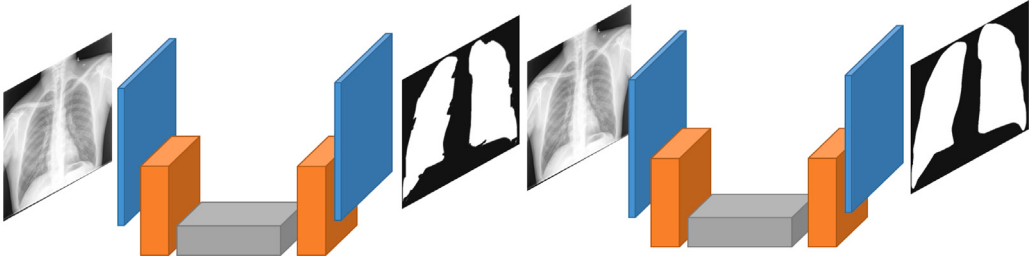


Fig. 3. Segmentation head in the architecture. The output of the first Unet is used as an attention map and concatenated with the input image for the input of the second Unet.

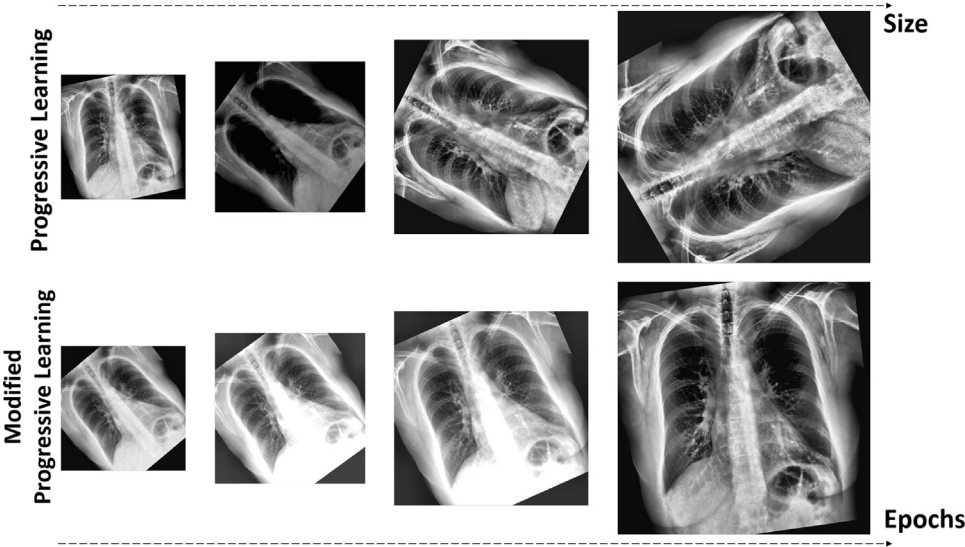


Fig. 4. Progressive learning vs amended progressive learning. The difference in the methodologies lies in the random augmentation factor which is kept constant for all sizes. Several augmentations can be applied to the same image. From left to right, the images show an increasing random augmentation factor with increasing input image size.

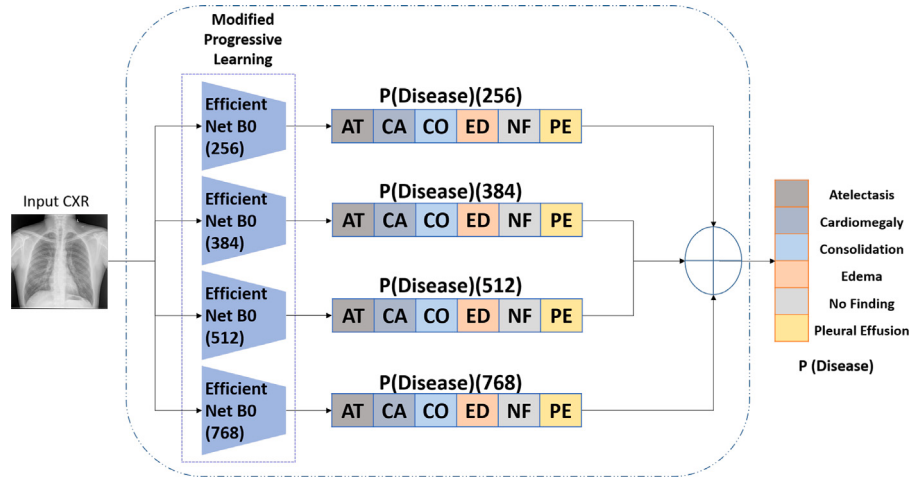


Fig. 5. Classification head used in the framework. The probability vector from each of the backbone is averaged to obtain the final $P(\text{Disease})$.

from BRAX [10]. Using the EfficientNet B0 as the feature extractor, Global Average Pooling was applied on the latent vector space. The final output layer for the model consisted of 6 output nodes with Softmax activation. As the same backbone architecture is used at different input sizes, therefore the weight migration from one trained model to the next is quite easy. Once all the models have been trained, they are combined in an ensemble with simple averaging to get the final disease probability $P(\text{Disease})$.

A number of augmentation techniques were used during the training strategy which included random rotation, flipping, zoom and shearing, random brightness, random sharpness, random blurring and CLAHE [55]. Area under the ROC curve (AUROC) was computed for a separate validation data set through which the performance of the model was determined and the best performing model was then saved. The learning rate strategy is kept the same as that for the segmentation training. For each backbone, the images were reshaped to the appropriate size but were not normalised.

3.2.3. Severity scoring

After the lungs have been segmented, opacities in the lung regions can then be localised. In the proposed framework, just like the classification head, an ensemble of multiple localisers is used that have been trained at different input resolutions of 512, 640 and 768 pixels using the Efficient Detector with D0, D1 and D2 [56]. The variation in the input resolution allows the ensemble to be able to detect opacities at different scales. In contrast to classification, the augmentation techniques used for localisers were limited to just horizontal flipping and rotation.

In order to improve the performance of the ensemble, two techniques were used: Test Time Augmentation (TTA) and Weighted Box Fusion (WBF) [57]. TTA performs the same augmentation techniques as those used at the training time and merges the result from a single architecture. WBF on the other hand, combines the outputs from different models in the ensemble to improve the overall accuracy. In order to gauge the performance of the model during the training, the mean average precision (mAP) score is used that is computed for the validation split and the best-performing model was kept just as for other heads.

4. Results

We evaluated the performance of our proposed framework on BRAX [10]. As [10] lacks a separate validation data set provided by the authors, so 5-fold cross-validation was performed for gauging the classification performance. For segmentation, performance evaluation

Table 3

F1 score for validation data sets for different segmentation model architectures.

Data sets	Model	Thresholding	Parameters (Millions)	F1 score	
				Validation	BRAX
[12,13]	U-Net	OTSU	34.5	0.9767	0.9162
	LU-net		0.032	0.9488	0.9114
	LW-net		0.068	0.9656	0.9246

was done on two sets: the validation data set of the three combined data sets and the manually annotated BRAX validation set.

As our severity scoring for different pathologies is based on the presence and prevalence of opacities in different lung regions, therefore, the results of opacity localisation and severity have been presented separately. For opacity localisation, SIIM validation data set is used. For severity scoring, we present our results on the BRAX validation data set.

For every model in the network, the model was trained for a minimum of three times and the best-performing model was kept. The models were trained using TensorFlow 2.8 in Python on a system with 64 GB RAM and two Nvidia RTX 2070 GPUs. In order to train some models on higher image resolution, we also made use of Google Cloud using Google TPUs (v2.8). The performance metrics that have been calculated for this framework include the AUROC score for classification, F1 score for segmentation, and mAP for opacity localisation.

4.1. Segmentation

Results from the validation data set and the BRAX data set are shown in Table 3. The F1 score difference between our suggested methodology and UNet on the validation data set is only 0.011, and it outperforms U-Net on the BRAX data set, where it achieves an F1 score of 0.9246.

In order to improve the performance on BRAX, we fine-tune our trained model on a set of 100 images. We start our fine-tuning by using merely 40 images and carry on increasing the number of images by 10 in every re-training of the model to determine the bare minimum number of images needed to outperform the earlier model. The performance of the model improves going from 0.9163 when 40 images are used to 0.9398 when all are used. Our fine-tuning strategy outperforms the previous model with just 50 images. Fig. 6 demonstrates that increasing the number of images for fine-tuning the model can have a positive effect on the performance of the model. In fact, from 6, it is clear that there is close to a linear relationship between performance and the quantity of images.

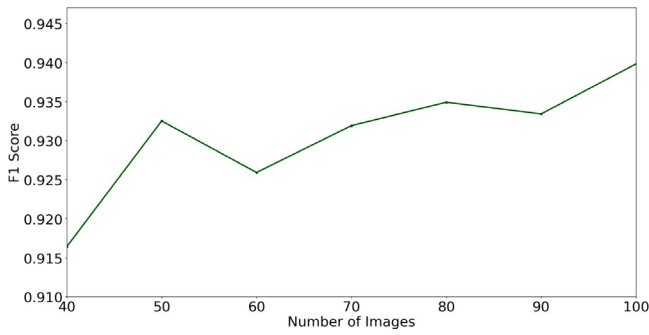


Fig. 6. F1 score increases with an increase in the number of fine-tuning images from the new data set.

Table 4

AUROC scores across 5 fold cross validation using amended progressive learning.

Fold, model	AUROC score						Avg. score
	AT	CA	CO	ED	NF	PE	
1, 256	0.795	0.889	0.777	0.829	0.867	0.906	0.844
1, 384	0.784	0.862	0.78	0.822	0.859	0.904	0.835
1, 512	0.763	0.891	0.79	0.817	0.852	0.919	0.839
1, 768	0.807	0.877	0.781	0.793	0.871	0.915	0.841
1, Ens	0.824	0.898	0.793	0.834	0.883	0.937	0.861
2, 256	0.79	0.914	0.803	0.94	0.872	0.896	0.869
2, 384	0.786	0.907	0.782	0.913	0.841	0.901	0.855
2, 512	0.779	0.896	0.797	0.941	0.879	0.913	0.868
2, 768	0.802	0.882	0.821	0.91	0.881	0.906	0.867
2, Ens	0.819	0.924	0.823	0.923	0.889	0.932	0.885
3, 256	0.763	0.851	0.806	0.833	0.837	0.869	0.826
3, 384	0.787	0.906	0.814	0.909	0.834	0.896	0.858
3, 512	0.784	0.893	0.815	0.928	0.852	0.907	0.863
3, 768	0.81	0.878	0.827	0.934	0.841	0.896	0.864
3, Ens	0.818	0.903	0.842	0.947	0.874	0.918	0.884
4, 256	0.788	0.888	0.827	0.857	0.867	0.908	0.856
4, 384	0.806	0.908	0.814	0.923	0.874	0.926	0.875
4, 512	0.789	0.901	0.844	0.917	0.868	0.897	0.869
4, 768	0.802	0.909	0.835	0.926	0.874	0.903	0.875
4, Ens	0.821	0.923	0.861	0.891	0.89	0.93	0.886
5, 256	0.755	0.892	0.796	0.942	0.863	0.896	0.857
5, 384	0.798	0.916	0.811	0.958	0.873	0.898	0.875
5, 512	0.789	0.903	0.819	0.901	0.882	0.896	0.865
5, 768	0.722	0.894	0.794	0.821	0.866	0.899	0.855
5, Ens	0.802	0.92	0.833	0.941	0.891	0.922	0.885

4.2. Classification

Although EfficientNet B0 [54] serves as the foundation of the proposed framework, we also assessed other networks (B1 to B7) in the same network family. Since, EfficientNet B0 model performed the best among the group, results presented here are based on this backbone. The performance on the data set is measured through AUROC. Using an ensemble of the 4 trained networks, the average AUROC achieved ranges from 0.861 to 0.886, which is higher than any single model at any size as the single best performing models only achieved an AUROC of 0.875 at two different sizes of 384 and 768 pixels. The worst-performing single model achieved an AUROC of 0.826 with an input size of 256 pixels. Furthermore, it can be observed from Table 4 that, on average, the best-performing class is Pleural Effusion while the worst-performing class is Atelectasis. Table 4 summarises the scores for each fold for all the models and the ensembles. The last column shows the average AUROC score.

Fig. 7 shows the ROC curves for all ensembles across folds. Even though the performance of Atelectasis is poor in comparison to the other classes is evident from its ROC curve, it still lies above the random classifier line. The small number of samples for Edema in each fold hinders the smoothness of its ROC curve.

The strength of the proposed framework lies in amended progressive learning. Using the weights from the preceding model trained at a

Table 5

AUROC scores across 5 fold cross validation using amended progressive learning with pre-training on CheXpert [9] data set.

Fold, model	AUROC score						Avg. score
	AT	CA	CO	ED	NF	PE	
1, 256	0.777	0.89	0.766	0.727	0.832	0.882	0.812
1, 384	0.727	0.889	0.766	0.837	0.83	0.86	0.818
1, 512	0.798	0.892	0.785	0.762	0.862	0.898	0.833
1, 768	0.784	0.898	0.79	0.772	0.875	0.883	0.834
1, Ens	0.802	0.925	0.804	0.813	0.882	0.912	0.856
2, 256	0.755	0.876	0.759	0.8	0.847	0.849	0.815
2, 384	0.765	0.914	0.774	0.835	0.87	0.905	0.844
2, 512	0.76	0.909	0.777	0.873	0.877	0.879	0.846
2, 768	0.757	0.879	0.774	0.851	0.842	0.893	0.833
2, Ens	0.801	0.926	0.804	0.921	0.88	0.924	0.876
3, 256	0.688	0.85	0.768	0.745	0.785	0.817	0.775
3, 384	0.759	0.887	0.808	0.91	0.839	0.893	0.849
3, 512	0.683	0.883	0.77	0.847	0.807	0.861	0.808
3, 768	0.748	0.892	0.819	0.84	0.828	0.874	0.834
3, Ens	0.767	0.915	0.828	0.878	0.85	0.891	0.855
4, 256	0.758	0.893	0.769	0.812	0.848	0.887	0.828
4, 384	0.775	0.912	0.822	0.828	0.884	0.931	0.859
4, 512	0.754	0.91	0.812	0.893	0.873	0.916	0.86
4, 768	0.752	0.9	0.81	0.907	0.854	0.907	0.855
4, Ens	0.785	0.931	0.826	0.907	0.883	0.941	0.879
5, 256	0.7	0.898	0.738	0.807	0.855	0.86	0.81
5, 384	0.771	0.911	0.763	0.815	0.868	0.887	0.836
5, 512	0.752	0.914	0.78	0.835	0.862	0.878	0.837
5, 768	0.764	0.903	0.77	0.916	0.868	0.879	0.85
5, Ens	0.773	0.933	0.785	0.878	0.887	0.91	0.861

Table 6

mAP score for SIIM [45] validation dataset for different models at 0.5.

Model	Image size	mAP
D0 [56]	512	0.4992
D1 [56]	640	0.4878
D2 [56]	768	0.5109
WBF [57]	–	0.5239

smaller input size, the next model is trained from these initial weights. It can be observed from Fig. 8 that the average AUROC for each class shows an improvement from the smallest input size to the largest where the maximum jump of 2.21% occurs for Atelectasis and the average jump for all the six classes is close to 1%. While Cardiomegaly shows improvement from the initial size of 256 pixels, the AUROC score for the largest size is the same as for the smallest size, therefore resulting in zero improvement.

Taking inspiration from [15], we also included a pre-training step that made use of the large-scale, publicly available data set Chexpert [9] with amended progressive learning. Instead of training from scratch, pre-trained Imagenet weights were used. Results of this training strategy are given in Table 5. With pre-training on [9], it can be seen that the average AUROC across 5 folds is within 1.5% of the results that do not include this step indicating that the pre-training does not have overly adverse effects.

4.3. Opacity localisation

The results for opacity localisation on SIIM [45] data set are shown in Table 6 which are reported using the PASCAL VOC at 0.5 threshold for mAP calculation. Efficient Detector [56] D0 to D2 have been used at varying image sizes of 512, 640, and 768. The results demonstrate that performance improved not only by increasing the network size but by also combining the results using Weighted Box Fusion. The weights for each classifier were determined empirically and are, for D0, D1, and D2, respectively, 0.2, 0.05, and 0.75. The scoring cut-off has been set for WBF [57] at 0.05 to maximise the score.

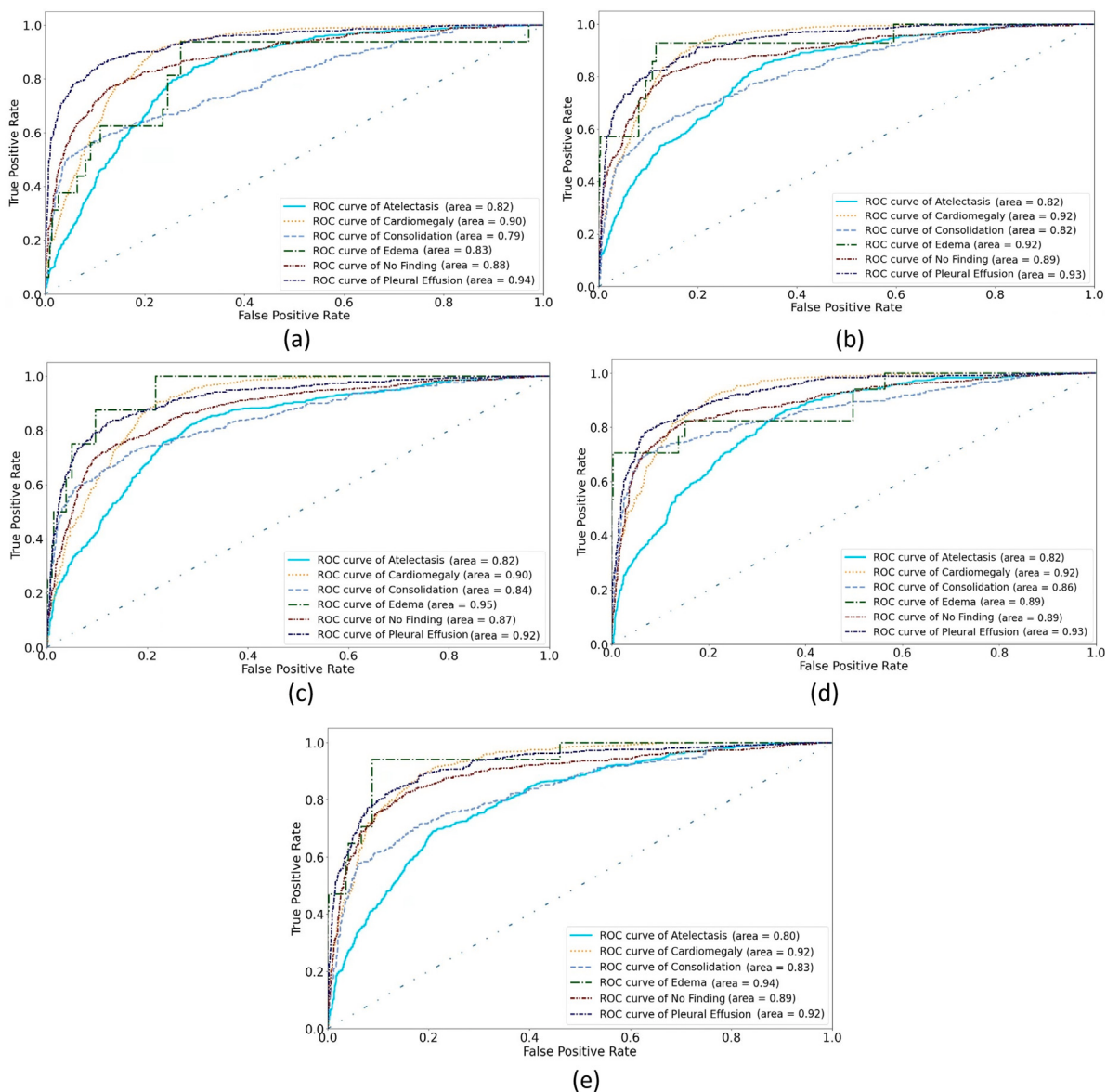


Fig. 7. AUROC curves for the ensembles across five folds ((a) to (e)). Ensembling the models together increase the performance for each fold.

4.4. Severity score

The presence or absence of opacity, denoted as 1 or 0, in each of the six lung regions is added to determine the severity score for each CXR and therefore the severity score for each CXR can vary between 0 to 6. A separate matching score that compares the performance of our framework to that of the radiologist in identifying the existence of opacity in a region for the entire CXR is calculated to assess its effectiveness. The greater the degree of agreement between the radiologist and the proposed framework on the presence or absence of opacity in each area, the higher this matching score will be which ranges from 0 to 1. Table 7 shows the mean matching score on the validation data set for different models at different confidence values at 0.3 Intersection over Union (IoU). The value of IoU was chosen empirically as it maximised the mean matching score.

Out of the 100 BRAX images that were manually annotated by the radiologist, the distribution of matching scores for these images can shed light on the performance of the severity scoring part of the framework. Fig. 9 provides insight into the performance of our suggested model. Fig. 9 shows that the framework receives a perfect score of 1 for 28 out of 100 images. It is off by just one region over

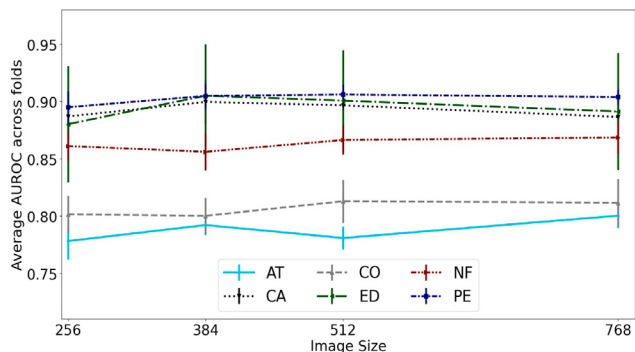


Fig. 8. Average AUROC for each class across different input image size shows the utility of progressive learning.

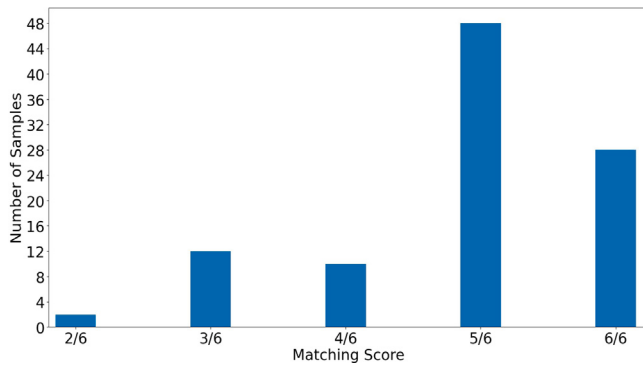


Fig. 9. Distribution of BRAX [10] validation set according to the matching score.

Table 7

Mean matching score opacities localised by different architectures at IoU value of 0.3.

Confidence	Efficient detector D0 [56]	Efficient detector D1 [56]	Efficient detector D2 [56]	WBF [57]
0.1	0.788	0.754	0.788	0.808
0.15	0.75	0.746	0.746	0.738
0.2	0.671	0.733	0.717	0.688
0.25	0.671	0.692	0.675	0.65
0.3	0.621	0.658	0.629	0.592
0.35	0.6	0.621	0.588	0.563
0.4	0.575	0.583	0.571	0.542
0.45	0.55	0.529	0.525	0.508
0.5	0.5	0.504	0.508	0.496
0.55	0.483	0.475	0.492	0.483
0.6	0.471	0.471	0.479	0.467
0.65	0.467	0.471	0.467	0.463
0.7	0.463	0.463	0.467	0.463

48 images. In contrast, there are only 2 images that have a difference of 4 regions' labels and 12 images that differ for three lung regions. The validation set's distribution reveals that the framework performs reasonably well, with a margin of error of just 1 region.

5. Discussion

The techniques presented here focus on several aspects, one of which is utilising a small number of images from an unknown target data set for segmentation. With as few as 50 images, we were able to demonstrate how this could enhance performance on the target data set. By fine-tuning the model using a small number of images, this technique can produce a generalised model whose performance can be continuously improved. For data sets where objects do not significantly differ between different data sets, this type of continuous training can be utilised as an alternative to [58]. Retraining for the new data set does not result in a significant time cost due to the network's small size.

In our experimentation, using backbones trained on large CXR data sets yielded lower performance than using Imagenet weights by 1.5%. However, some studies [14,15] have shown the potential of pre-training using large data sets. One of the reasons for this behaviour might be the small number of epochs that were used for training. Even though, the models were allowed to converge during training, increasing the number of epochs during this step might increase the performance in comparison to training directly from Imagenet weights. Another reason that may have resulted in this lower performance might be that even during pre-training, the weights were not randomly initialised but were initialised from Imagenet. For all our classification models, we used a softmax activation at the output layer while the performance metric was AUROC. In other studies [9,14,15,17], the

researchers opted for sigmoid activation. As softmax scales the probabilities such that they add up to 1, the decision to use this activation at the final layer might have resulted in some loss of performance as AUROC is concerned.

Basing the severity score of pathology on the opacities present in different lung zones enables the use of a single-digit score that can shed light on the progression of the disease. While the severity score is an indication of how many zones are affected by the disease, the matching score that has been proposed here can act as a confidence marker for the severity score. For the framework proposed here, it can be seen that for the BRAX validation set, for 28% of the images, the matching score is 1 i.e. the framework and the radiologist are in agreement for all the six regions of the lungs. However, if this threshold is dropped by just 1 region, then 75% of the images have the same opacity markings for both the framework and the human grader. This large jump can be explained by the nature of the opacities that may spill over from one region to the next and the IoU or the confidence or a combination of both may remove some of the opacities from the final scoring. In addition, the matching score can not only be used for gauging the performance of such frameworks but can even be extended to different human graders.

The framework proposed here achieves excellent performance for segmentation with a lightweight model. This performance can be improved even further by utilising more images from the BRAX [10] data set. The proposed modified progressive learning module has produced good classification results in form of AUROC. For severity scoring, a mean matching score of 80.8% shows that there is still room for improvement. Improving the opacity localisation by using diverse data sets can lead to an even better severity scoring performance.

Due to the fact that the framework we present here is made up of three distinct sub-modules, the overall computational complexity of the framework is an accumulation of the complexity of these modules. For opacity localisation, three different Efficient Detectors are used from D0 to D2 with 3.9 million, 6.6 million and 8.1 million parameters and 2.54 billion, 6.1 billion and 11 billion Floating Point Operations Per Second (FLOPS) [56]. Similarly, the classification head is composed of quadruple EfficientNet B0 at varying sizes which has 5.3 million parameters and 0.4 billion FLOPS at 224×224 size and scales up for the other sizes that have been used [54]. The segmentation head is the smallest one with only 0.068 million parameters. In essence, for each image at inference, more than 20 billion FLOPS are required.

The suggested framework's combination of progressive learning with a large data set results in one of its primary limitations. The training time and hardware requirements might quickly rise when using this methodology in conjunction with a huge data set because progressive learning necessitates training across a range of input sizes. The model's current training set only includes a portion of the pulmonary disorders that are contained in the data set, which is another drawback. The performance may suffer if the number of diseases for classification is increased. The proposed matching score is also currently region-based rather than pixel-based, which may offer a greater resolution for the disease's progression over time.

6. Conclusion

In this paper, we present a multi-head deep learning framework for pulmonary disease detection and severity scoring using opacity detection and localisation from a single chest X-ray and demonstrate reasonably good performance in all three components of the framework. The severity grading of pulmonary diseases helps to guide the subsequent processes as it can help determine what care should be provided to the patient based on the progression of the disease. We demonstrate the use of LW-Net for segmentation provides an efficient method for segmenting the lungs which can be continually improved for new target data sets by fine-tuning using a handful of images. Modified progressive learning which caps the augmentation rate of training

images for subsequent models in the pyramid has proved to be useful for model performance for classification on the BRAX data set. The collaboration with radiologists helped us in generating severity scores and to the best of our knowledge, the results of severity grading with the proposed techniques are the first of its kind along with the results for segmentation and classification. While our work aims to employ pre-training from a large scale data set, a future avenue that might be investigated is continual learning by training on other data sets to make the framework more adaptable and capable of handling X rays from varied populations. Lung segmentation performance, particularly for lungs affected by a disease to a great extent, can be improved. In addition, the number of pulmonary diseases covered in this work can also be increased in a continual fashion by making use of continual learning. The number of disorders for which a severity score can be awarded can also be expanded in the future along with increasing its resolution by providing a severity score for each affected pixel instead of a region.

CRedit authorship contribution statement

Asad Mansoor Khan: Data curation, Methodology, Software, Writing. **Muhammad Usman Akram:** Supervision, Conceptualization, Resources, Writing. **Sajid Nazir:** Validation, Original draft preparation, Writing. **Taimur Hassan:** Validation, Writing – review & editing. **Sajid Gul Khawaja:** Supervision, Validation, Writing – review & editing. **Tatheer Fatima:** Data curation, Resources.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

- [1] A.M. Tahir, M.E. Chowdhury, A. Khandakar, T. Rahman, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M.S. Rahman, S. Al-Maadeed, et al., COVID-19 infection localization and severity grading from chest X-ray images, *Comput. Biol. Med.* 139 (2021) 105002.
- [2] S. Manna, J. Wruble, S.Z. Maron, D. Toussie, N. Voutsinas, M. Finkelstein, M.A. Cedillo, J. Diamond, C. Eber, A. Jacobi, et al., COVID-19: a multimodality review of radiologic techniques, clinical utility, and imaging features, *Radiol. Cardiothorac. Imaging* 2 (3) (2020).
- [3] A. Narin, Accurate detection of covid-19 using deep features based on X-ray images and feature selection methods, *Comput. Biol. Med.* 137 (2021) 104771.
- [4] T. Dall, R. Reynolds, K. Jones, R. Chakrabarti, W. Iacobucci, The complexities of physician supply and demand: projections from 2017 to 2032, *Assoc. Am. Med. Coll.* (2019) 86.
- [5] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugupati, E.C. Rosenow III, Interpretation of plain chest roentgenogram, *Chest* 141 (2) (2012) 545–558.
- [6] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z.A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al., Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection, *Radiology* (2020).
- [7] E. Velichko, F. Shariaty, M. Orooji, V. Pavlov, T. Pervunina, S. Zavjalov, R. Khazaei, A.R. Radmard, Development of computer-aided model to differentiate COVID-19 from pulmonary edema in lung CT scan: EDECOVID-net, *Comput. Biol. Med.* 141 (2022) 105172.
- [8] L. Sun, Z. Wang, H. Pu, G. Yuan, L. Guo, T. Pu, Z. Peng, Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection, *Comput. Biol. Med.* 133 (2021) 104357.
- [9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, et al., Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 590–597.
- [10] E.P. Reis, J.P. de Paiva, M.C. da Silva, G.A. Ribeiro, V.F. Paiva, L. Bulgarelli, H.M. Lee, P.V. Santos, V.M. Brito, L.T. Amaral, et al., BRAX, Brazilian labeled chest X-ray dataset, *Sci. Data* 9 (1) (2022) 1–8.
- [11] A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, S. Horng, MIMIC-CXR-JPG-chest radiographs with structured labels (version 2.0. 0), *PhysioNet* (2019).
- [12] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, *Am. J. Roentgenol.* 174 (1) (2000) 71–74.
- [13] S. Jaeger, S. Candemir, S. Antani, Y.-X.J. Wang, P.-X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, *Quant. Imaging Med. Surg.* 4 (6) (2014) 475.
- [14] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, et al., BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset, *Med. Image Anal.* 71 (2021) 102046.
- [15] S. Park, G. Kim, Y. Oh, J.B. Seo, S.M. Lee, J.H. Kim, S. Moon, J.-K. Lim, J.C. Ye, Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification, *Med. Image Anal.* 75 (2022) 102299.
- [16] P.H. Gama, H. Oliveira, J.A. dos Santos, Learning to segment medical images from few-shot sparse labels, in: *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, IEEE*, 2021, pp. 89–96.
- [17] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexpert: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
- [18] W. Ye, J. Yao, H. Xue, Y. Li, Weakly supervised lesion localization with probabilistic-cam pooling, 2020, arXiv preprint arXiv:2005.14480.
- [19] A.S. Al-Waisy, S. Al-Fahdawi, M.A. Mohammed, K.H. Abdulkareem, S.A. Mostafa, M.S. Maashi, M. Arif, B. Garcia-Zapirain, COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images, *Soft Comput.* (2020) 1–16.
- [20] Y. Oh, J.C. Ye, Unifying domain adaptation and self-supervised learning for CXR segmentation via adain-based knowledge distillation, 2021, arXiv preprint arXiv:2104.05892.
- [21] M. Kim, B.-D. Lee, Automatic lung segmentation on chest X-rays using self-attention deep neural network, *Sensors* 21 (2) (2021) 369.
- [22] H. Azimi, J. Zhang, P. Xi, H. Asad, A. Ebadi, S. Tremblay, A. Wong, Improving classification model performance on chest X-rays through lung segmentation, 2022, arXiv preprint arXiv:2202.10971.
- [23] R. Jain, P. Nagrath, G. Kataria, V.S. Kaushik, D.J. Hemanth, Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning, *Measurement* 165 (2020) 108046.
- [24] Y. Cho, B. Park, S.M. Lee, K.H. Lee, J.B. Seo, N. Kim, Optimal number of strong labels for curriculum learning with convolutional neural network to classify pulmonary abnormalities in chest radiographs, *Comput. Biol. Med.* 136 (2021) 104750.
- [25] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2097–2106.
- [26] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Thorax disease classification with attention guided convolutional neural network, *Pattern Recognit. Lett.* 131 (2020) 38–45.
- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [28] Y. Oh, S. Park, J.C. Ye, Deep learning COVID-19 features on CXR using limited training data sets, *IEEE Trans. Med. Imaging* 39 (8) (2020) 2688–2700.
- [29] J.C. Souza, J.O.B. Diniz, J.L. Ferreira, G.L.F. da Silva, A.C. Silva, A.C. de Paiva, An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks, *Comput. Methods Programs Biomed.* 177 (2019) 285–296.
- [30] S. Zhao, P. Wang, A.A. Heidari, H. Chen, H. Turabieh, M. Mafarja, C. Li, Multilevel threshold image segmentation with diffusion association slime mould algorithm and Renyi's entropy for chronic obstructive pulmonary disease, *Comput. Biol. Med.* 134 (2021) 104427.
- [31] A.J. Selim, G. Fincke, X.S. Ren, W. Rogers, A. Lee, L. Kazis, A symptom-based measure of the severity of chronic lung disease: results from the Veterans Health Study, *Chest* 111 (6) (1997) 1607–1614.
- [32] H. Rodriguez, T.V. Hartert, T. Gebretsadik, K.N. Carroll, E.K. Larkin, A simple respiratory severity score that may be used in evaluation of acute respiratory infection, *BMC Res. Notes* 9 (1) (2016) 1–4.
- [33] J. Liu, W. Sun, X. Zhao, J. Zhao, Z. Jiang, Deep feature fusion classification network (DFFCNet): Towards accurate diagnosis of COVID-19 using chest X-rays images, *Biomed. Signal Process. Control* 76 (2022) 103677.
- [34] J. Zhu, B. Shen, A. Abbasi, M. Hoshmand-Kochi, H. Li, T.Q. Duong, Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs, *PLoS One* 15 (7) (2020) e0236621.
- [35] M. Tan, Q. Le, Efficientnetv2: Smaller models and faster training, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 10096–10106.

- [36] M.-L. Huang, Y.-C. Liao, A lightweight CNN-based network on COVID-19 detection using X-ray and CT images, *Comput. Biol. Med.* (2022) 105604.
- [37] A. Shoeibi, N. Ghassemi, J. Heras, M. Rezaei, J.M. Gorriz, Automatic diagnosis of myocarditis in cardiac magnetic images using CycleGAN and deep PreTrained models, in: *International Work-Conference on the Interplay Between Natural and Artificial Computation*, Springer, 2022, pp. 145–155.
- [38] A. Galdran, A. Anjos, J. Dolz, H. Chakor, H. Lombaert, I.B. Ayed, The little W-net that could: state-of-the-art retinal vessel segmentation with minimalistic models, 2020, arXiv preprint arXiv:2009.01907.
- [39] X. Xia, B. Kulis, W-net: A deep model for fully unsupervised image segmentation, 2017, arXiv preprint arXiv:1711.08506.
- [40] Y. Zhou, S.K. Wagner, M.A. Chia, A. Zhao, M. Xu, R. Struyven, D.C. Alexander, P.A. Keane, et al., AutoMorph: Automated retinal vascular morphology quantification via a deep learning pipeline, *Transl. Vis. Sci. Technol.* 11 (7) (2022) 12.
- [41] M.S. Kwak, J.M. Cha, J.W. Jeon, J.Y. Yoon, J.W. Park, Artificial intelligence-based measurement outperforms current methods for colorectal polyp size measurement, *Dig. Endosc.* (2022).
- [42] J.P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A.F. Abbasi, B. Shen, H.K. Mahsa, M. Ghassemi, H. Li, et al., Predicting covid-19 pneumonia severity on chest X-ray with deep learning, *Cureus* 12 (7) (2020).
- [43] H.Y.F. Wong, H.Y.S. Lam, A.H.-T. Fong, S.T. Leung, T.W.-Y. Chin, C.S.Y. Lo, M.M.-S. Lui, J.C.Y. Lee, K.W.-H. Chiu, T.W.-H. Chung, et al., Frequency and distribution of chest radiographic findings in patients positive for COVID-19, *Radiology* 296 (2) (2020) E72–E78.
- [44] E. Irmak, COVID-19 disease severity assessment using CNN model, *IET Image Process.* 15 (8) (2021) 1814–1824.
- [45] SIIM-FISABIO-RSNA, SIIM-FISABIO-RSNA COVID-19 Detection, 2021, <https://www.kaggle.com/c/siim-covid19-detection>. (Online accessed 26 August 2022).
- [46] L.R. Goodman, Felson's Principles of Chest Roentgenology, a Programmed Text, Elsevier Health Sciences, 2014.
- [47] M.A. Warren, Z. Zhao, T. Koyama, J.A. Bastarache, C.M. Shaver, M.W. Semler, T.W. Rice, M.A. Matthay, C.S. Calfee, L.B. Ware, Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS, *Thorax* 73 (9) (2018) 840–846.
- [48] M.A. Gedik, A. Keven, Radyoloji Başucu Serisi Pediatrik Radyoloji, Akademisyen Kitabevi, 2020.
- [49] A. Hattori, T. Matsunaga, K. Takamochi, S. Oh, K. Suzuki, Prognostic impact of a ground glass opacity component in the clinical T classification of non-small cell lung cancer, *J. Thorac. Cardiovasc. Surg.* 154 (6) (2017) 2102–2110.
- [50] E.A. Gaensler, C.B. Carrington, Peripheral opacities in chronic eosinophilic pneumonia: the photographic negative of pulmonary edema, *Am. J. Roentgenol.* 128 (1) (1977) 1–13.
- [51] V.S. Karkhanis, J.M. Joshi, Pleural effusion: diagnosis, treatment, and management, *Open Access Emerg. Med. OAEM* 4 (2012) 31.
- [52] D. Toussie, N. Voutsinas, M. Finkelstein, M.A. Cedillo, S. Manna, S.Z. Maron, A. Jacobi, M. Chung, A. Bernheim, C. Eber, et al., Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19, *Radiology* 297 (1) (2020) E197.
- [53] T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout, 2017, arXiv preprint arXiv:1708.04552.
- [54] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.
- [55] G. Yadav, S. Maheshwari, A. Agarwal, Contrast limited adaptive histogram equalization based enhancement for real time video system, in: *2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI*, IEEE, 2014, pp. 2392–2397.
- [56] M. Tan, R. Pang, Q.V. Le, Efficientdet: Scalable and efficient object detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10781–10790.
- [57] R. Solovyev, W. Wang, T. Gabruseva, Weighted boxes fusion: Ensembling boxes from different object detection models, *Image Vis. Comput.* 107 (2021) 104117.
- [58] M. Sirshar, T. Hassan, M.U. Akram, S.A. Khan, An incremental learning approach to automatically recognize pulmonary diseases from the multi-vendor chest radiographs, *Comput. Biol. Med.* 134 (2021) 104435.