EAI Endorsed Transactions
on Bioengineering and Bioinformatics
Research Article EAI.EU

Robust Heatmap Template Generation for COVID-19 Biomarker Detection

Mirtha Lucas[1,*], Miguel Lerma[2], Jacob Furst[3], Daniela Raicu[4]

[1]College of Computing and Digital Media, DePaul University, Chicago, United States, mlucas3@depaul.edu
[2]Department of Mathematics, Northwestern University, Evanston, United States, mlerma@math.northwestern.edu
[3]College of Computing and Digital Media, DePaul University, Chicago, United States, jfurst@cdm.depaul.edu
[4]College of Computing and Digital Media, DePaul University, Chicago, United States, draicu@cdm.depaul.edu

Abstract

INTRODUCTION: Detecting and identifying patterns in chest X-ray images of Covid-19 patients are important tasks for understanding the disease and for making differential diagnosis.
OBJECTIVES: The purpose of this work is to develop a technique for detecting biomarkers of four possible conditions in chest X-rays, and study the patterns arising from the location of biomarkers.
METHODS: We use transfer learning applied to a pretrained VGG19 neural network to build a model capable of detecting the four conditions in chest X-rays. For biomarkers detection we use Grad-CAM. Patterns in the biomarkers are found by using classical eigenfaces approach.
RESULTS: The discovered patterns are consistent across images from a given class of disease, and they are robust with respect to changes in dataset.
CONCLUSION: The identified patterns can serve as biomarkers for a given disease in chest X-ray images, and constitute explanations of how the deep learning model makes classification decisions.

Received on 16 December 2020; accepted on 21 February 2021; published on 24 February 2021

Keywords: Neural Networks, Biomarkers, Covid-19, Heatmaps

doi:10.4108/eai.24-2-2021.168729

1. Introduction

Covid-19 is a new acute disease that can be deadly, with an estimated 2% case fatality rate [20]. Early diagnosis may be beneficial for timely decisions about the course of action to take in each case. Medical imaging plays an important role in the process of detection and diagnosis. Computer-aided Diagnosis (CAD) systems may serve as a second opinion in complementing a physician's assessment [10].

Artificial Intelligence (AI) algorithms have shown great progress in pattern recognition tasks, and in particular for medical image analysis. During the last few years there has been a fast development of deep learning models for classification of images. These models have been embedded in state-of-the-art systems to detect Covid-19 from medical images, particularly chest X-rays. However, even these CAD systems present high prediction performance, many of them lack the transparency of showing how the results were produced and thus, they deepen the physicians' lack of trust in CAD [11]. Therefore, some kind of *explanation* of what the prediction is based on may allow the physicians to confirm, using their advanced domain knowledge, whether the prediction is likely to be correct. For example, for medical imaging, an explanation can come in the form of showing what area of the image has the largest impact in the outcome of the model.

Given that the Covid-19 pandemic appeared very recently, the available data from Covid-19 patients is limited compared to that of other diseases. A useful technique to develop models that work with small datasets is transfer learning. This technique consists of first training a model to classify samples from a large dataset. At the end of the initial training the model is assumed to have captured in its first layers the low-level features of the samples in the dataset, while high-level features leading to the final classification are captured

*Corresponding author. Email: mlucas3@depaul.com

in layers closer to its output. By freezing the first layers of the model and retraining only its last layers on the new, possibly smaller dataset, it is expected that the model will be able to capture the high-level features needed to perform classification of the samples of the new dataset.

Here we propose a transfer learning technique to develop a model able of detecting four possible conditions from chest X-ray images: normal (healthy), bacteria, virus (not Covid-19), and Covid-19. Furthermore, we work in the problem of *explainability*, i.e., how the model has arrived at the prediction. To that end we use the state-of-the-art Gradient Class Activation Map (Grad-CAM) technique described in [16] to identify the location of biomarkers, i.e, measurable indicators of the medical condition. Grad-CAM is able to determine which areas of an input image have the largest impact in each of the possible outputs of the network.

Grad-CAM and related techniques have been used extensively to locate which areas of an image contain some detected elements; for instance, in an image containing a dog and a cat, Grad-CAM is able to highlight the areas of the image where each of them appear. In the case of Chest X-rays used to detect a disease such as "Covid-19" the biomarkers may be clearly located (e.g. small lesions in an area of the lung), or may consist of general characteristics of the image that occupy large areas of the image (e.g. general transparency of the lung area). In this case the question remains to what extent the areas with the largest impact in the classification performed by the network depend on the particular image input, or whether those areas are relatively consistent across images of Chest X-rays for the same condition. We approach this question using principal component analysis and generate eigenheatmaps, the analogue of eigenfaces introduced by Turk et al. [19].

The rest of the paper is organized as follows. In Section 2 we discuss related work. Sections 3 and 4 present our methodology and results, respectively. Conclusions and future work are summarized in Section 5.

2. Related work

Computer-Aided Diagnosis (CAD) research has been successful in developing systems that can be used as second readers without increasing the demands on trained observers.

Although there is extensive work done on the algorithmic approaches for building the prediction models, often these CAD systems make a prediction without offering an explanation of how their predictions are made. Our hypothesis is that, when the diagnosis depends on interpretation of medical images, explanations can be provided by showing what areas of the image have the largest impact in the output of the CAD system. The expectations are that these regions might contain disease biomarkers that will lead to better medical image interpretation. Here we discuss recent research studies performed for detection of Covid-19 in Chest X-ray images (CXRs).

Apostolopoulos et al. (2020) [1] evaluated five different models, namely VGG19, Mobile Net, Inception, Xception, and Inception ResNet V2, pretrained with ImageNet followed with fine tuning, to predict if an image is from a COVID19 patient. While the accuracy of their final models for predicting Covid-19 was high (98.75% for VGG19, and 97.40% Mobile Net), the outcome of the network did not provide any explanations on what image content helped to determine one class versus another (e.g. bacteria versus Covid-19).

Basu et al. (2020) [2] acknowledged the limitations of transfer learning when the source and target domains are very dissimilar in nature, such as in natural images (like ImageNet) and medical images. Consequently, they used a large dataset of Chest X-ray images as the source domain. Overall their dataset consisted of 225 Covid-19 images from three open source databases, and 108,948 of Chest X-ray Images (not Covid-19). They arranged the samples into two datasets. Dataset A has two classes: normal, and disease. Dataset B has four classes: normal, other_disease, pneumonia, and Covid-19. Using a transfer learning technique, first training a convolutional neural network model was built from scratch on Dataset A. Next, they replaced and trained the last layer of their model to classify the four classes of Dataset B. The overall accuracy was measured as $95.3\% \pm 0.02$, with 100% of the Covid-19 and normal cases being correctly classified using 5-fold cross validation. There was some misclassification between pneumonia and other disease classes. Furthermore, the authors used Grad-CAM to detect the region where the model paid more attention during the classification.

Li et al. (2020) [14] went a step further by combining transfer learning and knowledge distillation [9] to produce a lightweight model that could be installed as a mobile application. After using a transfer learning technique, similar to the one used in [2], the authors used the retrained network, a DenseNet-121, to guide the training of a smaller network (MobileNetV2). In the last step MobileNetV2 was trained using a combination of loss functions based on the target predictions, and the KL-divergence with the soft outputs of the retrained DenseNet-121. The authors used further Grad-CAM to indicate that in fact their results show presence of Covid-19.

Karim et al (2020) [12] trained DenseNet-161, ResNet-18, and VGGNet-19 architectures several times each in a transfer learning setting, creating model snapshots. These architectures were incorporated into an ensemble, using Softmax class posterior averaging

and prediction maximization for the best performing models. Their dataset consisted of 16939 CXR images from 3 classes: normal (8066), pneumonia (8614), and Covid-19 (259). Heatmaps for all the test samples were generated based on the trained models, which indicated the relevance of each classification decision. Heatmaps are computed with three different techniques: Grad-CAM [16], Grad-CAM++ [6], and LPR (layer-wise relevance propagation) [4]. The heatmap depends on the model used to produce it and thus, the authors recommended to select the single best model as a basis to produce the heatmap.

Given the CAD research advances and the recent progress made in providing visual explanations based on techniques like Grad-CAM. we are proposing to analyze further the patterns present in the Grad-CAM heatmaps to determine if these patterns are consistent among CXR images of Covid-19 and thus, can be used to identify Covid-19 biomarkers.

3. Methodology

3.1. Dataset

We use the CoronaHack Chest X-ray dataset [8] that contains publicly available chest X-rays of Covid-19 patients. It consists of 5910 CXR images collected from various public sources, and divided into seven classes: normal (1576 images), bacteria (2772 images), virus different from Covid-19 (1493 images), Covid-19 (58 images), ARDS (2 images), SARS (4 images), Streptococcus (5 images). For this study, we removed the last three classes because they contained a very small number of images, so we worked with the following four classes only: normal, bacteria, virus, and Covid-19. Figure 1 shows a few sample images from the dataset. We will call it Dataset 1.

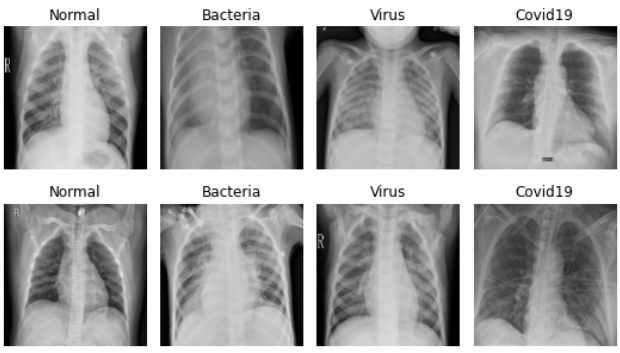

Figure 1. Sample CRXs

In order to study the robustness of the results we use a second dataset consisting of the same images used above plus additional Covid-19 X-ray images from the repository maintained by Cohen et al [7]. Not all images from that repository were adequate for use in our work. Some of them were CT-scans rather than X-ray images, and many contained extraneous elements such as annotations and medical instruments inside the image—some of the discarded images are shown in Fig. 2. After careful selection we managed to add to our dataset 116 new chest X-ray images of patients diagnosed with Covid-19. Dataset 1 plus the new 116 new chest X-ray images will be called Dataset 2.

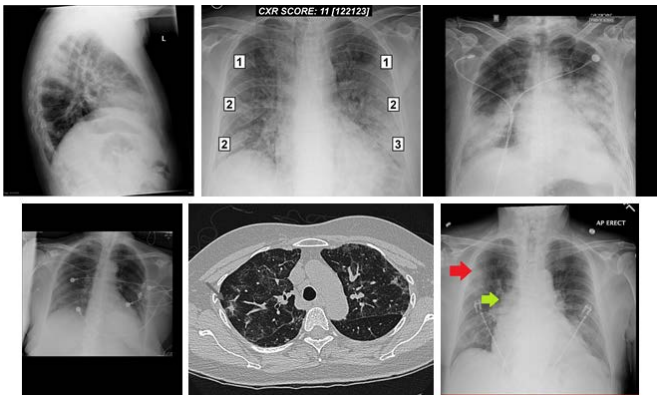

Figure 2. Some images discarded from the second dataset. The one on the center bottom is a CT-scan. The ones in center top and right bottom contain added annotation blocking parts of the image. The images in the top right and bottom left contained medical instruments.

3.2. Covid-19 Modeling

For the classification task, we use transfer learning [3]. To that end we use the pretrained VGG19 neural network trained on ImageNet [17]. This network contains 16 convolutional layers, 5 maxpooling layers, and 3 fully connected layers followed by softmax. The image input has size 224×224, and its final layer consists of 1,000 output units (Figure 4 (a)).

This model belongs to a collection of convolutional networks that have attained good performance in large scale image and video recognition, and can easily be imported for applications of transfer learning. The original VGG19 model can recognize one thousand classes of images, but our dataset has only four classes, so we decided to reduce the complexity of the top section of the network from three fully connected layers to just one global average layer followed by a single connected layer with only four outputs (one per class). Our training applies only to weights of the last convolutional layer and the added fully connected network (Figure 4 (b)). We also experimented with training on additional convolutional layers below the last one, but that did not have a significant impact in the final results. After training on Dataset 1 we call it Model 1.

For the robustness study we did first a preliminary experiment using our original Model 1 to generate

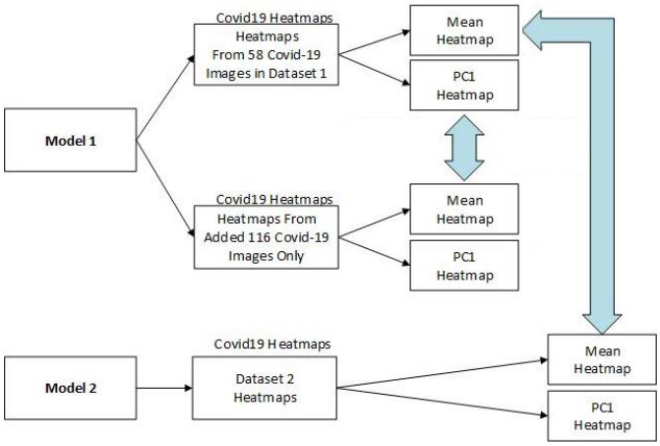

Figure 3. Top: Comparing the 58 Covid-19 heatmaps from Dataset 1 and the heatmaps of the added 116 Covid-19 images. Both sets of heatmaps produced with Model 1. Bottom: Comparing the Covid-19 heatmaps produced with Model 1 and Model 2.

heatmaps for Covid-19 images alone (Fig. 3, top). At this point we had two sets of heatmaps from Covid-19 images, let's call them H_{orig} and H_{new}. The set of heatmaps H_{orig} is generated using Model 1 on our original 58 Covid-19 images from Dataset 1, H_{new} is the set of heatmaps generated using Model 1 on the 116 added Covid-19 images. After performing PCA on H_{new} we got new average and principal components for the heatmap space. Then, we compared them to the ones we obtained for H_{orig}—the workflow is shown in to top part of Fig. 3.

The model obtained by training this network architecture with Dataset 2 will be called Model 2. One natural question was, what impact this change would have in the rest of the work, that is, the performance of the network as a classifier, and the heatmaps generated.

3.3. Biomarkers Detection

Biomarkers are measurable indicators of a medical condition. In the case of lung conditions, we are interested in visible lung lesions. For biomarkers detection, we use the state-of-the-art Grad-CAM technique [16]. This technique computes the gradients of the activation function of each class output with respect to the activations of the last convolutional layer, and combine the results to obtain a coarse heatmap with information on the relevance of each area of the input image that contributes to the output of the selected class output of the network. More specifically, given a convolutional layer (typically the last one) of size $Z = u \times v$ (here u and v represent the first two dimensions of the layer, i.e., height and width respectively, the third dimension is its number of *channels* or *feature maps*—these two terms are synonyms in this context) in a neural network, and a class c, the Grad-CAM technique consists of computing the gradient of the score for y^c (before the softmax) for that class with respect to the activations A^k_{ij} of the feature maps of the convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k_{ij}}$ (computed via the backpropagation algorithm). Here k indexes the feature map (channel) of the chosen convolutional layer, and i, j vary along the width and height dimensions of the layer. The gradients are global-averaged-pooled over the width and height dimensions to obtain the importance weights α^c_k of each channel of the chosen convolutional layer:

$$\alpha^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{A^k_{ij}}}_{\text{gradients via backprop}} \tag{1}$$

Then a weighted combination of forward activation maps is performed and a rectified linear unit (ReLU) [15] is applied to the result to take into account only the features that have a positive effect on the class of interest:

$$L^c_{\text{Grad-CAM}} = ReLU\left(\sum_k \alpha^c_k A^k\right) \tag{2}$$

The $u \times v$ matrix $L^c_{\text{Grad-CAM}}$ obtained can be interpreted as a coarse heatmap of the same size as the convolutional layer picked. The heatmap can be resized to the size of the input image for comparison; Figure 5 shows an example of a CXR image and its combination with its heatmap obtained using the Grad-CAM technique.

3.4. Average Heatmaps, and Eigen-Heatmaps

After generating Grad-CAM heatmaps for various images of CXRs we observe that there are areas of the image that consistently have larger impact for a given class output, for instance the upper central area of the image has a larger contribution to an output of "normal" than to "bacteria", "virus", or "Covid-19". Our hypothesis is that in fact the network pays special attention to certain areas when deciding whether a sample image belongs to a given class.

In order to determine the areas of an image that most contribute to each class output we start by computing the average heatmap for each class along the images that belong to a given class. That way, for each class we obtain average heatmaps for the four classes (Figure 6). This produces a total of $4 \times 4 = 16$ average heatmaps corresponding to all possible pairs: {image class, heatmap class}. We expect to see that the average heatmap corresponding to each class (say the average heatmap of "virus" computed for the images in the "virus" class) will have pixel values larger than the

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv4 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv4 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv4 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0

flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000

Total params: 143,667,240
Trainable params: 143,667,240
Non-trainable params: 0

(a) VGG19 Network

global_average_pooling2d (Gl	(None, 512)	0
dense (Dense)	(None, 4)	2052

Total params: 20,026,436
Trainable params: 2,361,860
Non-trainable params: 17,664,576

(b) Our Network

Figure 4. VGG19 and Our Network

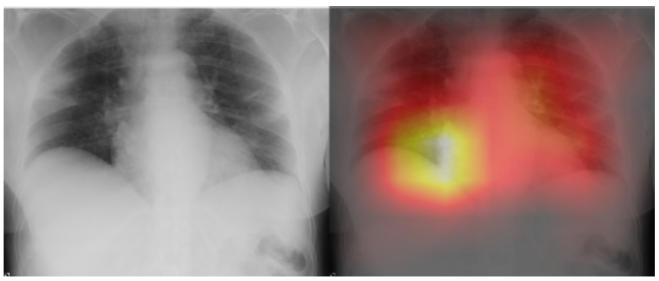

Figure 5. Sample original X-ray and overlaid heatmap

heatmaps computed for the other classes (heatmaps of "normal", "bacteria" and "Covid-19" for images within the "virus" class).

Next, interpreting each heatmap as a vector in a $224 \times 224 = 50176$ dimensional space, we perform a PCA analysis to determine the variability of the heatmaps for images of each class (say heatmaps of "virus" within the "virus" class), a technique that mimics the "eigenfaces" technique used by Turk et al. (1991) for face recognition [19]. The main difference with the eingefaces technique is that our heatmaps are not images, but raw heatmaps as computed by the Grad-CAM algorithm before they are normalized using min-max normalization. The algorithm always produces non-negative pixel values for the (raw) heatmaps due to the use of the ReLU activation function, but in principle there is no upper limit for the possible pixel values computed.

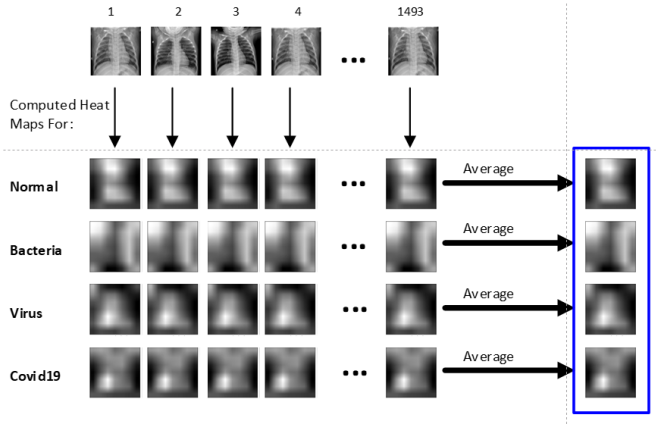

Figure 6. Illustration of how average heatmaps are calculated for each class of images—virus in this case (1493 images). For each image in the class we use Grad-CAM to generate heatmaps for each of the four classes (normal, bacteria, virus, Covid-19). Then we average the heatmaps generated for each class.

4. Experiments and Results

We show first the results obtained using Model 1 on Dataset 1. Then, in a final subsection about robustness, we proceed with the results obtained using Model 1 on Dataset 2, and finally the results of using Model 2 on Dataset 2.

4.1. Training

For training the network we split the dataset into 80% training and 20% validation using stratified sampling as shown in Table 1.

Table 1. Dataset Split

	Training	Validation	Total
normal	1260	316	1576
bacteria	2217	555	2772
virus	1194	299	1493
Covid-19	46	12	58

The loss function used was sparse categorical cross-entropy, and the training algorithm was RMSprop [18] with a learning rate of 0.0001 for 20 epochs. The progress of the training is shown in Figure 7. We observed that the learning metrics improved slowly, but using longer training led to overfitting and decided to stop it at the 20 epochs mark.

4.2. Multiclass Classifier

After finishing the training of the network, we tested it as a multiclass classifier. Its performance was measured using a confusion matrix (Table 2), classification report (Table 3), and "one vs the rest" receiver

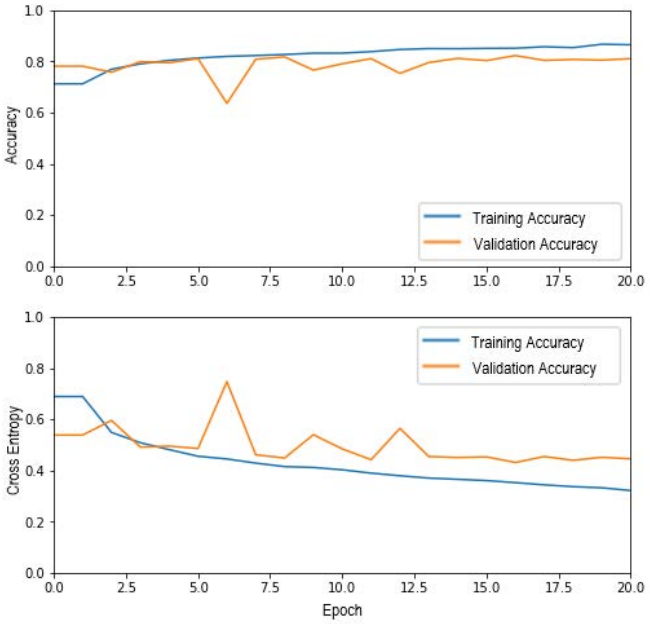

Figure 7. The top plot shows the training and validation accuracy. The bottom plot shows the training and validation cross entropy loss.

operator characteristic curves (ROC) (Figure 8). In the classification report, *precision*, *recall*, and *F1-score* are reported. The *macro average* is the arithmetic mean of the values of a statistical measure across the classes; in Table 3 the macro average was calculated for precision, recall and F1-score. The *weighted average* is the same but weighted by the number of elements of each class; similarly, the weighted average was calculated for precision, recall and F1-score. The *support* is the number of elements in each class.

Table 2. Confusion matrix of classifier

		Predicted			
		normal	**bacteria**	**virus**	**Covid-19**
Actual	**normal**	299	0	17	0
	bacteria	7	436	112	0
	virus	10	75	214	0
	Covid-19	2	0	0	10

The results are reasonably good, particularly for prediction of Covid-19, with an area under the curve (AUC) almost 1. Given the good performance in detecting Covid-19 we decided to test the use of the network as a binary classifier, as explained next.

4.3. Binary Classifier for Covid-19

Based on the Covid-19 ROC curve we can design a binary classifier to detect the presence or absence of the

Table 3. Classification report

	precision	recall	f1-score	support
normal	0.94	0.95	0.94	316
bacteria	0.85	0.79	0.82	555
virus	0.62	0.72	0.67	299
Covid-19	1.00	0.83	0.91	12
accuracy			0.81	1182
macro avg	0.85	0.82	0.83	1182
weighted avg	0.82	0.81	0.81	1182

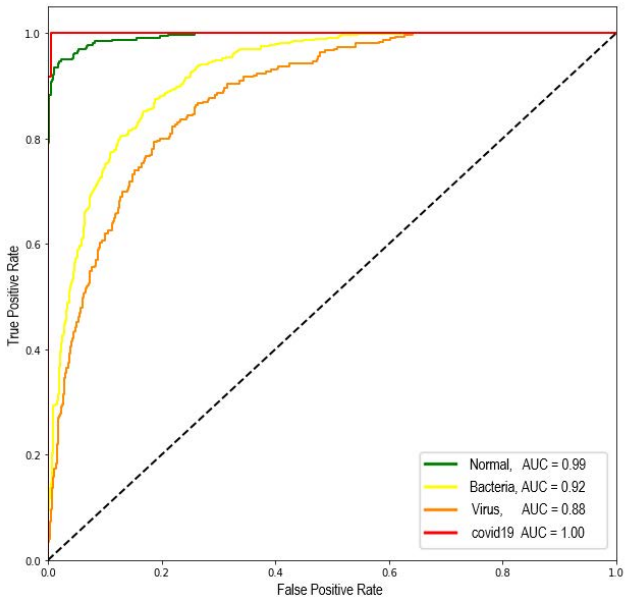

Figure 8. One versus all Receiving Operating Curves for normal, bacteria, virus, and Covid-19. Note the large values of the area under the curve for all the classes.

disease using a threshold for the predicted probability. In our model, we found that a threshold of 0.051 was the best one producing no false negatives and a minimum amount of false positives. In other words, if p_{pred} is the probability of Covid-19 predicted by the network, then our binary classifier will classify CXR image as follows:

$$\text{Covid-19?} = \begin{cases} \text{YES} & \text{if } p_{pred} \geq threshold \\ \text{NO} & \text{if } p_{pred} < threshold \end{cases} \quad (3)$$

The threshold may vary if the network is retrained, but in all of our experiments the optimal threshold ended up being very close to the one shown here.

The confusion matrix and classification report on the test set are in Tables 4 and 5 respectively.

4.4. Biomarkers Location - Grad-CAM Heatmaps.

Next, we apply the Grad-CAM technique to locate areas of each CXR image that have relevance in the output of each class. The convolutional layer picked

Table 4. Confusion matrix of Covid-19 binary classifier

	pred. negatives	pred. positives
actual negatives	1164	6
actual positives	0	12

Table 5. Classification report for Covid-19 binary classifier

	precision	recall	f1-score	support
not Covid-19	1.00	0.99	1.00	1170
Covid-19	0.67	1.00	0.80	12
accuracy			0.99	1182
macro avg	0.83	1.00	0.90	1182
weighted avg	1.00	0.99	1.00	1182

for building the heatmap is the last 7×7 maxpooling (*block5_pool*), right before the last global average and final dense layer. Figure 9 shows an example of CXR image with its heatmap generated using the Grad-CAM technique. The original image is shown to the left, and its combination with a heatmap to the right. To make it more visible a colormap has been used for the heatmap, with lower values represented with darker tones, medium values are red, and larger values are yellow to white. In the discussion that follows, the heatmaps will be represented in grayscale for simplicity.

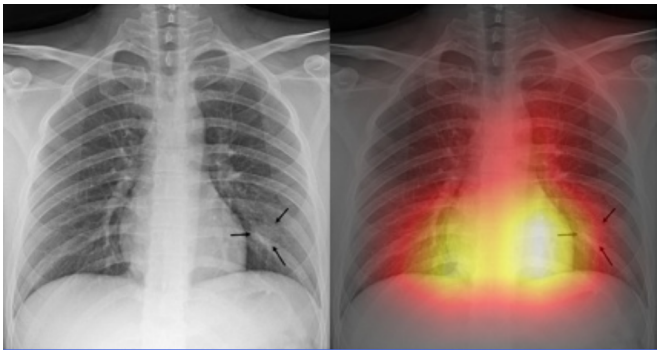

Figure 9. Original X-ray and overlaid heatmap

The Grad-CAM technique generates different heatmaps for different images, but we found them to be relatively consistent along images of each class. By averaging the values of the raw heatmap across each class we get the results shown in Figure 10.

We observe that 'virus' and 'Covid-19' have similar heatmaps, showing that the area with largest impact in the output of the network is located on the left bottom area of the image, although for Covid-19 the maximum is a little higher up, and also it has some area around the upper right corner not present in the heatmap for virus. The heatmap for 'bacteria' covers mainly two vertical almost symmetrically placed bands at both sides of the image. The biomarkers for 'normal' are shown as

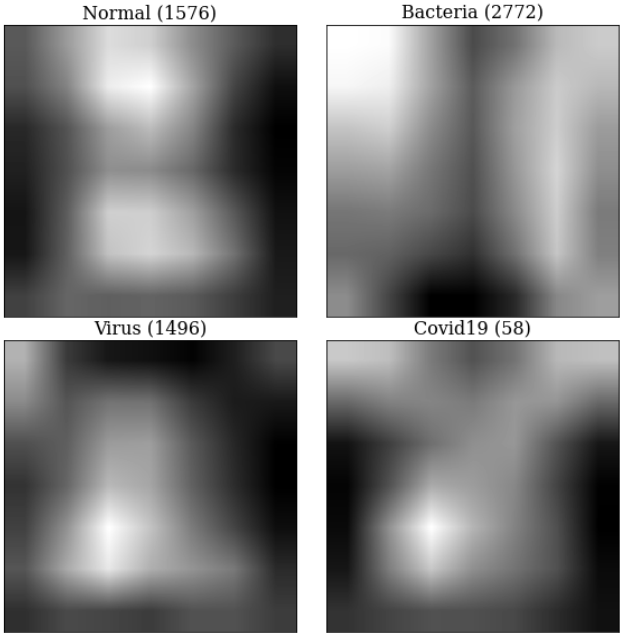

Figure 10. Average Heatmaps for each class. In parenthesis the number of elements in each class.

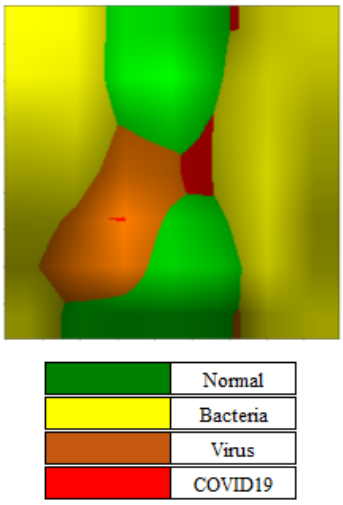

Figure 11. Four color representation of the model response for normal, bacteria, virus, and Covid-19. The image is a combination of the average heatmaps, after min-max normalization, and assigning to each pixel the color corresponding to the class with maximum intensity at that pixel.

covering two approximately horizontal bands on top and bottom of the heatmap. Given that we expect the biomarker to coincide with some sort of lung lesion, in the case of 'normal', i.e., no disease, the heatmap should have darker areas in the places where lesions are present in case of disease. While this is somehow true when comparing the heatmap for normal with the one for bacteria, we see noticeable overlapping in the lower area of the heatmaps for normal, virus and Covid-19. Note however that the last convolutional layer of the network contains a large number of channels (512 channels to be precise), each capturing some different feature from the image, and it is perfectly conceivable that although different, those features may partially overlap in the image.

Note that in order to represent the heatmaps as images we had to adjust their grayscales so that the minimum value represented 'black', and the maximum was 'white' (min-max normalization). Consequently, there is some information loss in the graphic representation given by Figure 10. In order to provide some additional information about the distribution of actual (raw) values for the average heatmap of each class we plotted their whisker-plots the five-number summary of the pixel value distributions (maximum, minimum, median, and the upper and lower quartiles), as shown in Figure 12. We observe that for each class, the whisker-plot of the average heatmap of that class dominates over the rest. This is particularly evident in the boxplot for Covid-19, where the whisker-plot for Covid-19 shows larger values than the corresponding whisker-plots for normal, bacteria, and virus. Only in the class 'virus' we see an overlapping of the whisker-plots for 'virus' and bacteria'.

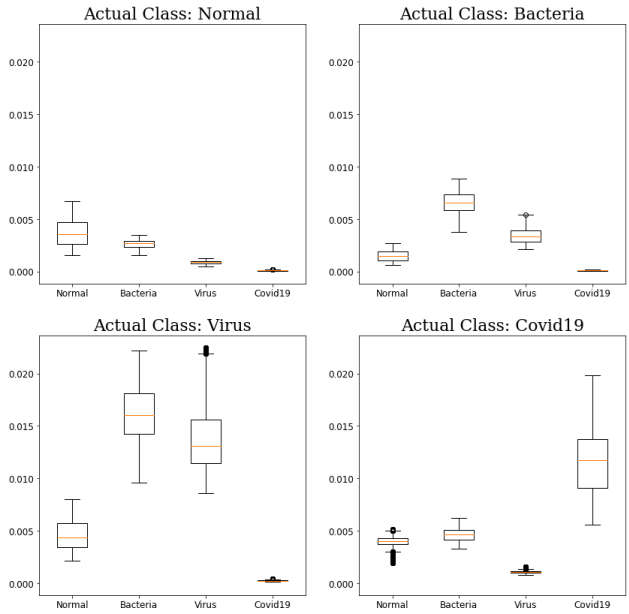

Figure 12. Box plots distributions of pixel values of average heatmaps for each class.

4.5. Eigen-Heatmaps

Although the average heatmaps provide potentially useful information about what areas of an image have the largest impact in the network predictions for each class, at this point we do not yet have much information about how the heatmaps vary across each class of CXR images. Our next step is to answer this question by using a technique inspired in Turk at al's *eigenfaces* [19]. The idea consists of interpreting each heatmap as a real vector in a linear space of dimension equal to the number of pixels ($224 \times 224 = 50176$ in our case), and perform dimensionality reduction using PCA.

For each class, we apply PCA analysis and retain the principal components that explain 95% of the variance. Figure 13 shows (as images) the first three principal components for class Covid-19. Note how similar the first principal component is to the average heatmap for Covid-19. The same happens with the other classes too, i.e., the first principal component of each class is very similar to the average heatmap for that class. We discuss this below.

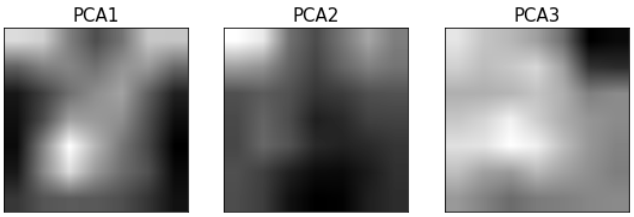

Figure 13. First three PCA components for class Covid-19.

Table 6. Variance explained by each principal component

	class			
principal component	normal	bacteria	virus	Covid-19
PC1	0.92	0.91	0.78	0.88
PC2	0.02	0.02	0.05	0.03
PC3	0.01	0.01	0.03	0.02
PC4			0.03	0.02
PC5			0.02	0.01

Figure 14 shows an example of reconstruction of a Covid-19 heatmap using the average heatmap and the first principal components. Original heatmap is at the center top. The images in the 2×3 grid from left to right and top to bottom are the average heatmap, and the successive reconstructions of the original heatmap. The image at the top left is the average heatmap for Covid-19. The next image is the average heatmap plus a multiple of the first principal component (PC1), and so on. We see how the image on the right bottom closely resembles the original heatmap.

We notice that the dimensionality reduction is very sharp—see proportion of variance explained by each principal component in Table 6. Furthermore, just the first principal component explains most of the variance, with very little contribution from any of the following components.

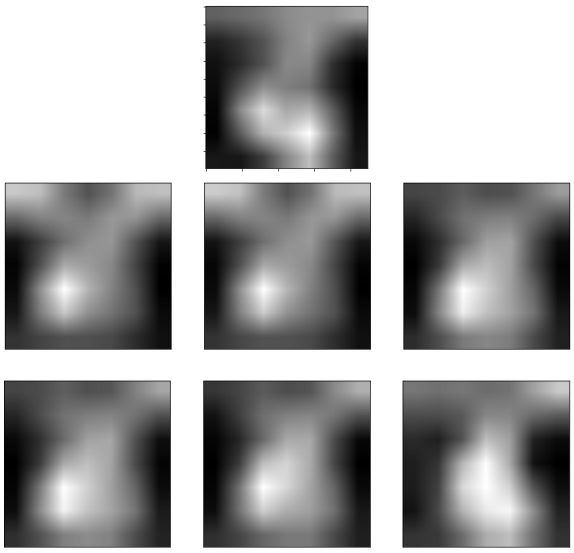

Figure 14. Heatmap reconstruction of a Covid-19 heatmap using the average heatmap and the first principal components. Original heatmap is at the center top. The images in the 2×3 grid from left to right and top to bottom are the average heatmap, and the successive reconstructions of the original heatmap. The image at the top left is the average heatmap for Covid-19. The next image is the average heatmap plus a multiple of the first principal component (PC1), and so on. We see how the image on the right bottom closely resembles the original heatmap.

Table 7. Correlation between average heatmaps and first principal component of each class

	normal AVG	normal PC1	bacteria AVG	bacteria PC1
normal AVG	**1.00**	**0.99**	-0.31	-0.26
normal PC1	**0.99**	**1.00**	-0.30	-0.26
bacteria AVG	-0.31	-0.30	**1.00**	**0.95**
bacteria PC1	-0.26	-0.25	**0.95**	**1.00**
virus AVG	0.52	0.55	-0.26	-0.12
virus PC1	0.35	0.39	-0.18	-0.04
Covid-19 AVG	0.66	0.68	0.034	0.08
Covid-19 PC1	0.63	0.64	0.036	0.08
	virus AVG	**virus PC1**	**Covid-19 AVG**	**Covid-19 PC1**
normal AVG	0.52	0.35	0.66	0.63
normal PC1	0.55	0.39	0.68	0.64
bacteria AVG	-0.26	-0.18	0.034	0.04
bacteria PC1	-0.12	-0.04	0.08	0.08
virus AVG	**1.00**	**0.97**	0.58	0.55
virus PC1	**0.97**	**1.00**	0.53	0.50
Covid-19 AVG	0.58	0.53	**1.00**	**0.99**
Covid-19 PC1	0.55	0.50	**0.99**	**1.00**

Also, we note that the Pearson correlation of each average heatmap with respect to its first PCA component is close to 1 (see Table 7, correlations between each average heatmap and first principal component are in bold).

Note that in this study we are using *raw* heatmaps as computed by the Grad-CAM algorithm without the min-max normalization that will allow later to show them as images. The fact that the first principal component has such high correlation with the average heatmap is an indication that the main source of variability in the heatmaps across each class is approximately given by an affine transformation of the average heatmap, i.e., changing the scale of the pixel values of the (raw) average heatmap and adding a constant. Only after the min-max normalization used to show them as images their small differences become more pronounced.

4.6. Robustness study

As indicated in the methodology section, in our first experiment we used our original Model 1 to generate heatmaps for all Covid-19 images. The results of the correlation analysis of those heatmaps are shown in Table 8. The high correlations indicate that the heatmaps generated by Grad-CAM on our Model 1, using new never seen Covid-19 images, are very similar to the ones obtained for the images that we used to train the network. For instance, the correlation between average heatmaps of H_{orig} and H_{new} is 0.97.

Table 8. Correlation between average heatmaps and first principal component of Covid-19 images, original and new dataset

	original AVG	original PC1	new AVG	new PC1
original AVG	1.00	0.99	0.97	0.97
original PC1	0.99	1.00	0.96	0.95
new AVG	0.97	0.96	1.00	0.99
new PC1	0.97	0.95	0.99	1.00

In the robustness study, the Dataset 2 used for training and validation contains the additional 116 Covid-19 images, split as shown in Table 9.

Table 9. New Dataset Split

	Training	Validation	Total
normal	1260	316	1576
bacteria	2217	555	2772
virus	1194	299	1493
Covid-19	139	35	174

For training Model 2 we used the same hyperparameters as in the original Model 1. Working as a multiclass classifier the results were very similar to those obtained for Model 1, as shown in the new confusion matrix (Table 10), classification report (Table 11), and ROC curves (Fig. 15).

Following the steps we performed with the original Model 1, we generate heatmaps for all images, and the average heatmaps across the elements of each class are shown in Fig. 16.

Table 10. New confusion matrix of classifier

		Predicted			
		normal	**bacteria**	**virus**	**Covid-19**
Actual	**normal**	293	4	18	1
	bacteria	11	411	133	0
	virus	13	53	234	0
	Covid-19	1	0	3	31

Table 11. New classification report

	precision	recall	f1-score	support
normal	0.92	0.93	0.92	316
bacteria	0.88	0.74	0.80	555
virus	0.60	0.78	0.68	300
Covid-19	0.97	0.87	0.93	35
accuracy			0.80	1206
macro avg	0.84	0.83	0.83	1206
weighted avg	0.82	0.80	0.81	1206

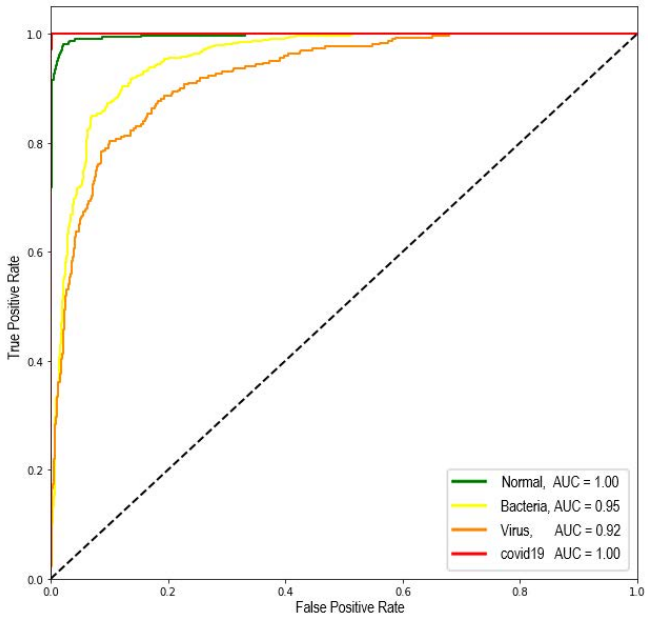

Figure 15. New one versus all Receiving Operating Curves for normal, bacteria, virus, and Covid-19.

The whisker-plots of the average heatmaps produced with Model 2 are also shown in Fig. 17. The whisker-plots are very similar to the ones obtained with the Model 1, except for the class "virus," where the whisker-plots for virus now appears above the one for bacteria, which is consistent with the desired classification ability of the model.

The Pearson correlation between average heatmaps and first PCA components for each class (Table. 12) are also similar to the results obtained using Model 1,

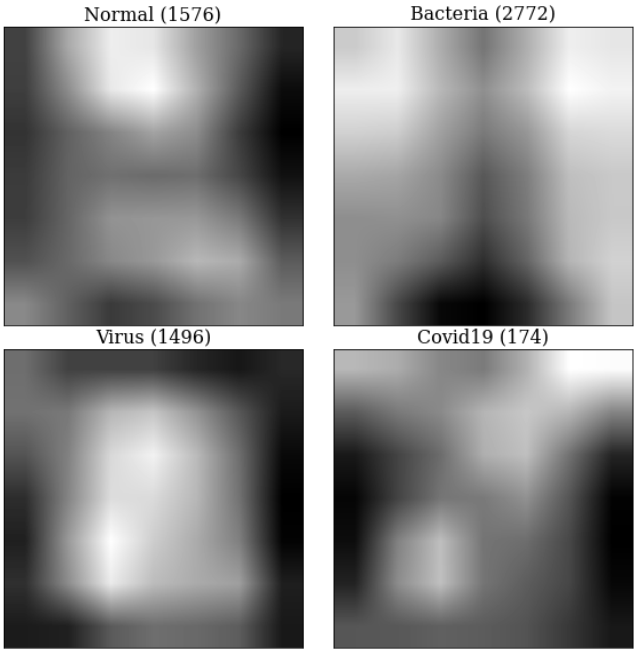

Figure 16. New Average Heatmaps for each class. In parenthesis the number of elements in each class.

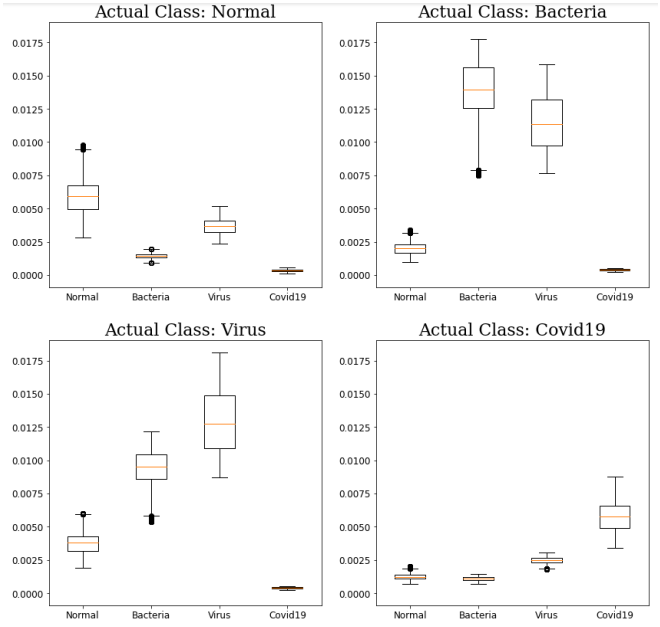

Figure 17. Box plots distributions of pixel values of average heatmaps for each class.

so each PC1 is still highly correlated to the average heatmap of the same class.

Finally, we computed the correlations between the average heatmaps and PC1 components of Covid-19 images obtained using Model 1 and Model 2. This provides information about how much the heatmaps change after changing the model. As shown in Table 13, the correlations are still high (larger than 0.85), hence the heatmaps for Covid-19 images do not change much after modifying the model.

Table 12. New correlations between average heatmaps and first principal component of each class

	normal AVG	normal PC1	bacteria AVG	bacteria PC1
normal AVG	**1.00**	**0.99**	-0.25	-0.20
normal PC1	**0.99**	**1.00**	-0.25	-0.20
bacteria AVG	-0.25	-0.25	**1.00**	**0.96**
bacteria PC1	-0.20	-0.20	**0.95**	**1.00**
virus AVG	0.47	0.48	-0.43	-0.26
virus PC1	0.19	0.21	-0.29	-0.08
Covid-19 AVG	0.36	0.30	0.055	0.08
Covid-19 PC1	0.26	0.19	0.016	0.04
	virus AVG	**virus PC1**	**Covid-19 AVG**	**Covid-19 PC1**
normal AVG	0.47	0.19	0.36	0.26
normal PC1	0.48	0.21	0.30	0.19
bacteria AVG	-0.43	-0.29	0.055	0.017
bacteria PC1	-0.27	-0.08	0.082	0.04
virus AVG	**1.00**	**0.89**	0.36	0.30
virus PC1	**0.89**	**1.00**	0.19	0.15
Covid-19 AVG	0.36	0.19	**1.00**	**0.98**
Covid-19 PC1	0.30	0.15	**0.98**	**1.00**

Table 13. Correlation between average heatmaps and first principal component of Covid-19 images, original and new model

	original AVG	original PC1	new AVG	new PC1
original AVG	1.00	0.99	0.89	0.85
original PC1	0.99	1.00	0.89	0.87
new AVG	0.89	0.89	1.00	0.98
new PC1	0.85	0.87	0.98	1.00

5. Conclusions and future work

We have used a transfer learning technique to develop a neural network capable of detecting four different conditions from chest X-ray images, one of them the novel Covid-19. The results are good despite the small size of the dataset used for training. The Covid-19 condition is particularly well detected, and the multiclass classifier can easily be transformed into a binary classifier capable of detecting Covid-19 with almost 100% accuracy. Furthermore, we used the state-of-the-art Grad-CAM technique for the location of biomarkers of the conditions in the X-ray images. A significant contribution of this work is the observation that there are some regions of the CRX images that have a larger impact on the classification of each of the medical conditions. Furthermore, those areas are relatively consistent across the images of each class. Their consistency in location and appearance makes them good candidates for becoming "templates" indicating where and for what structures to pay special attention when looking for a particular disease.

Given that one of the conditions, namely Covid-19, is new, the available data is still limited compared to the other conditions. By adding new images of Covid-19 X-rays and slightly modifying the architecture of the model, we showed that the heatmap patterns

obtained remain stable, but more work may be needed as additional data becomes available.

We are aware that the heatmaps were generated at the level of the last layer, of size 7×7. Therefore, we believe heatmaps generated using layers further below the output of the network would have higher resolution, and may reveal finer details in the area that contributes to the network output.

References

[1] Ioannis D. Apostolopoulos, Tzani Bessiana (2020). Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 43, 635–640 (2020).

[2] Sanhita Basu, Sushmita Mitra, Nilanjan Saha (2020). Deep Learning for Screening COVID-19 using Chest X-ray Images, arXiv preprint arXiv:2004.10507 [eess.IV]

[3] Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. Proceedings of ICML workshop on unsupervised and transfer learning, 2012, pp. 17–36.

[4] Alexander Binder, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, Wojciech Samek (2016). Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. International Conference on Artificial Neural Networks, 2012, pp 63–71.

[5] William C. Black, and H. Gilbert Welch. Advances in Diagnostic Imaging and Overestimations of Disease Prevalence and the Benefits of Therapy. April 29, 1993 N Engl J Med 1993; 328:1237–1243.

[6] Aditya Chattopadhay and Anirban Sarkar. 2018. Grad-CAM++: Generalizedgradient-based visual explanations for convolutional networks. In Applications of Computer Vision(WACV). IEEE, 839–847.

[7] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, Marzyeh Ghassemi (2020) COVID-19 Image Data Collection: Prospective Predictions Are the Future. arXiv preprint arXiv:2006.11988 [q-bio.QM]

[8] Praveen Govindaraj (2020). CoronaHack Chest X-ray Dataset. [Online; accessed 31-March-2020] https://www.kaggle.com/praveengovi/coronahack-chest-xraydataset

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean (2015). Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning Workshop, 2015.

[10] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. Artificial intelligence in radiology. Nat Rev Cancer. 2018 Aug; 18(8): 500–510.

[11] W. Jorritsma, F. Cnossen, and P.M.A. van Ooijen. Improving the radiologist–CAD interaction: designing for appropriate trust. Clinical Radiology, Volume 70, Issue 2, P115–122, February 01, 2015 https://www.overleaf.com/project/5efcf4cf7a2a600001f07f97

[12] Md. Rezaul Karim, Till Döhmen, Dietrich Rebholz-Schuhmann, Stefan Decker, Michael Cochez, Oya Beyan (2020). DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images. arXiv preprint arXiv:2004.04582 [eess.IV]

[13] Kelly, C.J., Karthikesalingam, A., Suleyman, M. et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17, 195 (2019).

[14] Xin Li, Chengyin Li, Dongxiao Zhu (2020). COVID-MobileXpert: On-Device COVID-19 Screening using Snapshots of Chest X-ray (2020), arXiv preprint arXiv:2004.03042 [eess.IV]

[15] Vinod Nair and Geoffrey Hinton (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning June 2010 Pages 807-–814.

[16] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[17] Rebecca Smith-Bindman, Diana L. Miglioretti, and Eric B. Larson. Rising Use Of Diagnostic Medical Imaging In A Large Integrated Health System. Health Aff (Millwood). 2008 Nov-–Dec; 27(6): 1491—1502.

[18] Ruder, Sebastian (2017). An overview of gradient descent optimization algorithms. Insight Centre for Data Analytics, NUI GalwayAylien Ltd., Dublin.

[19] Matthew A Turk, Alex P Pentland (1991). Face recognition using eigenfaces. Computer Vision and Pattern Recognition, Proceedings CVPR'91., IEEE Computer Society Conference on 1991.

[20] Z. Zu, L. Shi, Y. Wang, et al. (2020). Pathological findings of COVID-19 associated with acute respiratory distress syndrome. The Lancet Respiratory Medicine.