COVIDGR dataset and COVID-SDNet methodology for predicting COVID-19 based on Chest X-Ray images

S. Tabik[a,1], A. Gómez-Ríos[a], J.L. Martín-Rodríguez[b], I. Sevillano-García[a], M. Rey-Area[c], D. Charte[a], E. Guirado[d], J.L. Suárez[a], J. Luengo[a], M.A. Valero-González[b], P. García-Villanova[b], E. Olmedo-Sánchez[b], F. Herrera[a]

[a] *Andalusian Research Institute in Data Science and Computational Intelligence University of Granada, 18071, Spain*
[b] *Hospital Universitario Clnico San Cecilio de Granada, Spain*
[c] *atlanTTic Research Center for Telecommunication Technologies, University of Vigo, Galicia, Spain*
[d] *Multidisciplinary Institute for Environment Studies Ramn Margalef, University of Alicante, 03690, Spain*

Abstract

Currently, Coronavirus disease (COVID-19), one of the most infectious diseases in the 21st century, is diagnosed using RT-PCR testing, CT scans and/or Chest X-Ray (CXR) images. CT (Computed Tomography) scanners and RT-PCR testing are not available in most medical centers and hence in many cases CXR images become the most time/cost effective tool for assisting clinicians in making decisions. Deep learning neural networks have a great potential for building triage systems for detecting COVID-19 patients, especially patients with low severity. Unfortunately, current databases do not allow building such systems as they are highly heterogeneous and biased towards severe cases. This paper is three-fold: (i) we demystify the high sensitivities achieved by most recent COVID-19 classification models, (ii) under a close collaboration with Hospital Universitario Clnico San Cecilio, Granada, Spain, we built COVIDGR-1.0, a homogeneous and balanced database that includes all levels of severity, from Normal with positive RT-PCR, Mild, Moderate to Severe. COVIDGR-1.0 contains 377 positive and 377 negative PA (PosteroAnterior) CXR views and (iii) we propose COVID Smart Data based Network (COVID-SDNet) methodology for improving the generalization capacity of COVID-classification models. Our approach reaches good and stable results with an accuracy of 97.37% ± 1.86%, 88.14%±2.02%, 66.5%±8.04% in severe, moderate and mild COVID severity levels. Our approach could help in the early detection of COVID-19. COVIDGR-1.0 dataset will be made available after the review process.

Keywords: COVID; Smart Data; Convolutional neural networks

[*] Corresponding author

Preprint submitted to Elsevier *June 3, 2020*

1. Introduction

In the last months, the world has been witnessing how COVID-19 pandemic is increasingly infecting a large mass of people very fast everywhere in the world. The trends are not clear yet but some research confirm that this problem may persist until 2024 [1]. Besides, prevalence studies conducted in several countries reveal that a tiny proportion of the population have developed antibodies after exposure to the virus, e.g., 5% in Spain [1]. This means that frequently a large number of patients will need to be assessed in small time intervals by few number of clinicians and with very few resources.

In general, COVID-19 diagnosis is carried out using at least one of these three tests.

- Computed Tomography (CT) scans-based assessment: it consists in analyzing 3D radiographic images from different angles. The needed equipment for this assessment is not available in most hospitals and it takes more than 15 minutes per patient in addition to the time required for CT decontamination [2].
- Reverse Transcription Polymerase Chain Reaction (RT-PCR) test: it detects the viral RNA from sputum or nasopharyngeal swab [3]. It requires specific material and equipment, which are not easily accessible and it takes at least 12 hours, which is not desirable as positive COVID-19 patients should be identified and tracked as soon as possible. Some studies found that RT-PCR results from several tests at different points from the same patients were variable during the course of the illness producing a high false-negative rate [4]. The authors suggested that RT-PCR test should be combined with other clinical tests such as CT.
- Chest X-Ray (CXR): The required equipment for this assessment are less cumbersome and can be lightweight and transportable. In general this type of resources is more available than the required for RT-PCR and CT-scan tests. In addition, CXR test takes about 15 seconds per patient [3]. Which makes CXR one of the most time/cost effective assessment tools.

Few recent studies provide estimates on expert radiologists sensitivity in the diagnosis of COVID-19 based on CT scans, RT-PCR and CXR. A study on a set of 51 patients with chest CT and RT-PCR essay performed within 3 days, reported a sensitivity in CT of 98% compared with RT-PCR sensitivity of 71% [5]. A different study on 64 patients (26 men, mean age 56 ± 19 years) reported a sensitivity of 69% for CXR compared with 91% for initial RT-PCR [3]. According to an analysis of 636 ambulatory patients [6], most patients presenting to urgent care centers with confirmed coronavirus disease 2019 have

[1] https://english.elpais.com/society/2020-05-14/antibody-study-shows-just-5-of-spaniards-have-contracted-the-coronavirus.html

normal or mildly abnormal findings on CXR. Only 58.3% of these patients are correctly diagnosed by the expert eye.

In a recent study [3], authors proposed simplifying the quantification of the level of severity by adapting a previously defined Radiographic Assessment of Lung Edema (RALE) score [7] to COVID-19. This new score is calculated by assigning a value between 0-4 to each lung depending on the extent of visual features such as, consolidation and ground glass opacities, in the four parts of each lung as depicted in Figure 1. Based on this score, experts can identify the level of severity of the infection among four severity stages, Normal 0, Mild 1-2, Moderate 3-5 and Severe 6-8. In practice, a patient classified by expert radiologist as Normal can have positive RT-PCR. We refer to these cases as Normal-PCR+. Expert annotation adopted in this work is based in this score.

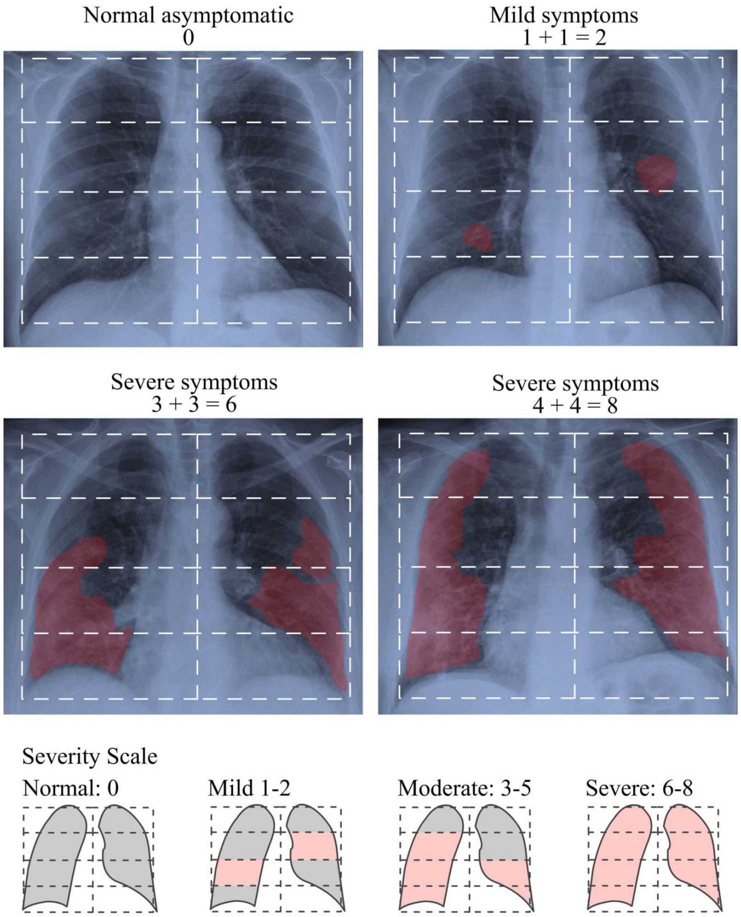


Figure 1: The stratification of radiological severity of COVID. Examples of how RALE index is calculated.

Automated image analysis via Deep learning (DL) models have a great potential to optimize the role of CXR images for a fast diagnosis of COVID-19. A robust and accurate DL model could serve as a triage method and as a support for medical decision making. An increasing number of recent works claim achieving impressive sensitivities $> 95\%$, far higher than expert radiologists. These high sensitivities are due to the bias in the most used COVID dataset, *COVID-19 Image Data Collection* [8]. This dataset includes a very small number of COVID positive cases, coming from highly heterogeneous sources (at least 15 countries) and most cases are severe patients, an issue that drastically reduces its clinical value. To populate Non-COVID and Healthy classes, AI researchers are using CXR images from diverse pulmonary disease repositories. The obtained models will have no clinical value as well since they will be unable to detect patients with low and moderate severity, which are the target of a clinical triage system. In view of this situation, there is still a huge need for higher quality datasets built under the same clinical protocol and under a close collaboration with expert radiologists.

The concept of Smart Data refers to the process of converting raw data into higher quality data with higher concentration of useful information [9]. Multiple studies have proven that higher quality data ensures higher quality models. Smart data includes all pre-processing methods that improve value and veracity of data. Examples of these methods include noise elimination, data-augmentation [10] and data transformation [11] among other techniques.

In this work, we designed a high clinical quality dataset, named COVIDGR-1.0 that includes four levels of severity, Normal-PCR+, Mild, Moderate and Severe. We identified these four severity levels from a recent COVID radiological study [3]. We also propose COVID Smart Data based Network (COVID-SDNet) methodology. It combines segmentation, data-augmentation and data transformations together with an appropriate Convolutional Neural Network (CNN) for inference.

The contributions of this paper can be summarized as follows:

- To analyze reliability, potential and limitations of the most used COVID CXR datasets and models.
- To provide a high quality dataset, called COVIDGR-1.0, for building triage systems with high clinical value.
- To design a novel methodology, named COVID-SDNet, with a high generalization capacity for COVID classification based on CXR images. COVIDSDNet combines segmentation, data-transformation to increase the discrimination capacity of the classification model, data-augmentation, and a suitable CNN model together with an inference approach to get the final class.

Experiments demonstrate that our approach reaches good and stable results especially in moderate and severe levels, with $97.37\% \pm 1.86\%$ and $88.14\% \pm 2.02\%$ respectively. Lower accuracies were obtained in mild and normal-PCR+ severity levels with $66.5\% \pm 8.04\%$ and $38.68\% \pm 2.44\%$ respectively.

This paper is organized as follows: A review of the most used datasets and COVID classification approaches is provided in Section 2. Section 3 describes how COVIDGR-1.0 is built and organized. Our approach is presented in Section 4. Experiments, comparisons and results are provided in Section 5 and finally Conclusions are pointed out in Section 6.

2. Related works

The last three months have known an increasing number of works exploring the potential of deep learning models for automating COVID-19 diagnosis based on CXR images. The results are promising but still too much work needs to be done at the level of data and models design. Given the potential bias in this type of problems, several studies include explication methods to their models. This section analyzes the advantages and limitations of current datasets an models for building automatic COVID-19 diagnosis systems with and without decision explication.

2.1. Datasets

There does not exist yet a high quality collection of CXR images for building COVID diagnosis systems of high clinical value. Currently, the main source for COVID class is *COVID-19 Image Data Collection* [8]. It contains 76 positive and 26 negative PA views. These images were obtained from highly heterogeneous equipment from all around the world. To build Non-COVID classes, most studies are using CXR from one or multiple public pulmonary disease data-sets. Examples of these repositories are:

- RSNA Pneumonia CXR challenge dataset on Kaggle [12].
- Figure-1-COVID- 19 Chest X-ray Dataset Initiative [13]
- ChestX-ray8 dataset [14].
- MIMIC-CXR dataset [15].
- PadChest dataset [16].

For instance, COVIDx 1.0 [17] was built by combining three public datasets: (i) *COVID-19 Image Data Collection* [8], (ii) Figure-1-COVID- 19 Chest X-ray Dataset Initiative [13] and (iii) RSNA Pneumonia Detection Challenge dataset [12]. COVIDx 2.0 was built by re-organizing COVIDx 1.0 into three classes, Normal (healthy), Pneumonia and COVID-19 using 201 CXR images for COVID class, including PA(PosteroAnterior) and AP(AnteroPosterior) views (see Table 1). Notice that for a correct learning front view (PA) and back view (AP) cannot be mixed in the same class.

Although the value of these datasets is unquestionable as they are being useful for carrying out first studies and reformulations, however they do not guarantee useful triage systems for the next reasons. It is not clear what annotation protocol has been followed for constructing the positive class in *COVID-19*

Version	Normal(healthy)	Pneumonia	COVID-19
1.0	1,583	4,273 (Bacterial+viral)	76
2.0	8,066	8,614	190

Table 1: A brief description of COVIDx dataset [8] (only PA views are counted).

Image Data Collection. The included data is highly heterogeneous and hence DL-models can rely on other aspects then COVID visual features to differentiate between the involved classes. This dataset does not provide a representative spectrum of COVID-19 severity levels, most positive cases are of severe patients [18].

Our claim is that the design of a high quality dataset must be done under a close collaboration between expert radiologists and AI experts. The annotations must follow the same protocol and representative numbers of all levels of severity, especially Mild and Moderate levels, must be included.

2.2. DL classification models

Existing related works are not directly comparable as they consider different combinations of public data-sets and different experimental setup. A brief summary of these works is provided in Table 2.

Ref.	Classes	Datasets	Model	Partition	Sens.	Acc.
[17]	Normal, Pneumonia, COVID	COVIDx 1.0	COVIDNet	98% - 2%	87.1%	92.6%
[19]	Normal, COVID	COVIDx 1.0	COVID-CAPS	98% - 2%	90%	95.7%
[20]	No-Findings, COVID No-Findings, Pneumonia, COVID	[8] + [14]	DarkCovidNet	5-FCV 5-FCV	90.65% 97.9%	98.08% 87.02%
[21]	Normal, Pneumonia, COVID	COVIDx 2.0+[12]	VGG-19 + DenseNet-161	70% - 30%	93%	96.77%
[22]	Normal, Bacterial, Viral, COVID	[8]+[12]	Bayesian ResNet50V2	80% - 20%	85.71%	89.82%
[23]	Normal, Pneumonia, COVID	[8] + [12] + other sources	MobileNet	10-FCV	98.66%	96.78%

Table 2: Summary of related works that analyze variations of COVIDx with CNN.

The most related studies to ours as they proposed different models to the typical ones are [17] and [19]. In [17], the authors designed a deep network, called COVIDNet. They affirmed that COVIDNet reaches an overall accuracy of 92.6%, with 97.0% sensitivity in Normal class, 90.0% in Non-COVID-19 and 87.1% in COVID-19. The authors of a smaller network, called COVID-CAPS [19], also claim that their model achieved an accuracy of 98.7%, sensitivity of 90%, specificity of 95.8%. These results look too impressive when compared to expert radiologist sensitivity, 69%. This can be explained by the fact that the used dataset is biased to severe COVID cases [18]. In addition, the performed experiments in both cited works are not statistically reliable as they were evaluated on one single partition. The stability of these models, in terms of standard deviation, has not been reported.

DL classification models with explanation approaches: Several interesting explanations were proposed to help inspect the predictions of DL-models [22, 21] although all their classification models were trained and validated on variations of COVIDx. The authors in [21] first use an ensemble of two CNN networks to predict the class of the input image, as Normal, Pneumonia or COVID. Then highlight class-discriminating regions in the input CXR image using gradient-guided class activation maps (Grad-CAM++) and layer-wise relevance propagation (LRP). In [22], the authors proposed explaining the decision of the classification model to radiologists using different saliency map types together with uncertainty estimations (i.e., how certain is the model in the prediction).

3. COVIDGR 1.0: Data acquisition, annotation and organization

It is well known that the larger is the database the more effective is the learning of ML algorithms. Even when the data is of lower quality, algorithms can actually perform better, as long as useful information can be extracted by the model. Alternatively, instead of starting with an extremely large and noisy dataset, one can build a small and smart dataset then augment it in a way it increases the performance of the model. This approach has proven effective in multiple studies. This is particularly true in the medical field, where access to data is heavily protected due to privacy concerns and costly expert annotation.

Under a close collaboration with four highly trained radiologists from Hospital Universitario Clnico San Cecilio, Granada, Spain, we first established a protocol on how CXR images are selected and annotated to be included in the dataset. A CXR image is annotated as COVID-19 positive if both RT-PCR test and expert radiologist confirm that decision within less than 24 hours. CXR with positive PCR are labeled as Normal-PCR+. The involved radiologists annotated the level of severity of positive cases based on RALE score as: Normal-PCR+, Mild, Moderate and Severe. Patients with positive RT-PCR that were annotated by expert radiologists as Normal are actually asymptomatic patients.

Dataset	Class	#images	women	men	#img. per severity level
COVIDGR-1.0	Negative	377	211	166	
	COVID-19	377	164	213	Normal-PCR+: 76 Mild: 80 Moderate: 145 Severe: 76

Table 3: A brief summary of COVIDGR-1.0 dataset. All samples in COVIDGR 1.0 are segmented CXR images considering only PA view.

COVIDGR-1.0 is organized into two classes, positive and negative. It contains 754 images distributed into 377 positive and 377 negative cases, more details are provided in Table 3. All the images were obtained from the same equipment and under the same X-ray regime. Only PosteriorAnterior (PA) view is considered. COVIDGR-1.0 will be available to the scientific community after review at `https://github.com/ari-dasci/covidgr`.

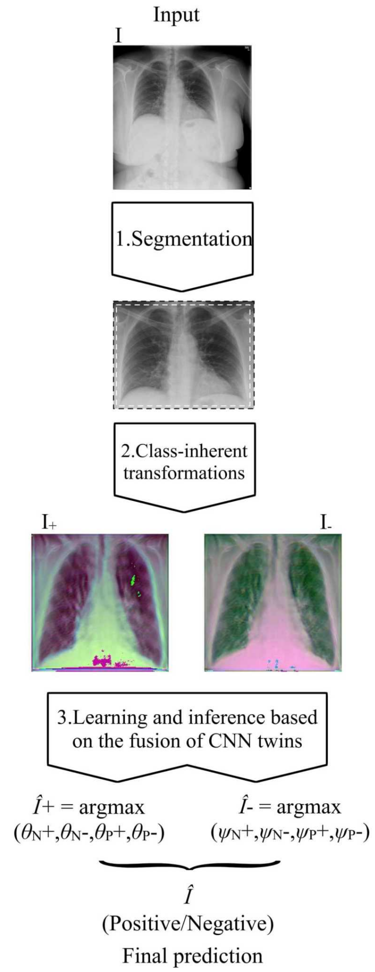

Input
I
1.Segmentation
2.Class-inherent transformations
I_+
I_-
3.Learning and inference based on the fusion of CNN twins
$\hat{I}+ = \text{argmax} (\theta_N+, \theta_N-, \theta_P+, \theta_P-)$
$\hat{I}- = \text{argmax} (\psi_N+, \psi_N-, \psi_P+, \psi_P-)$
$\hat{I}$
(Positive/Negative)
Final prediction

Figure 2: Flowchart of the proposed COVID-SDNet methodology.

4. COVID-SDNet methodology

In this section, we describe COVID-SDNet methodology in detail, covering pre-processing to produce smart data, including segmentation and data transformation for increasing discrimination between positive and negative classes, combined with a deep CNN for classification.

One of the pieces of COVID-SDNet is the CNN-based classifier. We have selected Resnet-50 initialized with ImageNet weights for a transfer learning approach. To adapt this CNN to our problem, we have removed the last layer of the net and added 512 neurons layer with ReLU activation and two or four neuron layer (according to the considered number of classes) with softmax activation. All the layers of the network were fine-tuned. We used a batch size of

16 and SGD as optimizer.

The main stages of COVID-SDNet are three, two associated to pre-processing for producing quality data (smart data stages) and the learning and inference process. A flowchart of COVID-SDNet is depicted in Figure 2.

1. *Segmentation: Unnecessary information elimination*
   Different CXR equipment brands include different extra information about the patient in the sides and contour of CXR images. The position and size of the patient may also imply the inclusion of more parts of the body, e.g., arms, neck, stomach. As this information may alter the learning of the classification model, first, we used the pre-trained U-Net segmentation model provided in [24] to first extract the smallest rectangle that includes left and right lungs. Then, to avoid eliminating useful information, we add 2.5% of pixels to the left, right, up and down sides of the rectangle. An illustration with example of this pre-processing is shown in Figure 3.

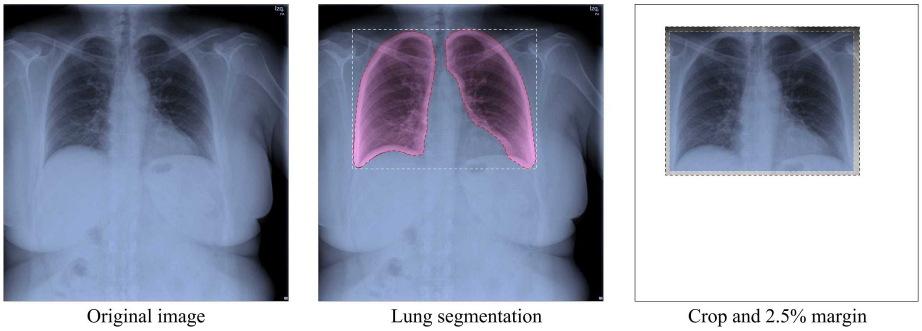
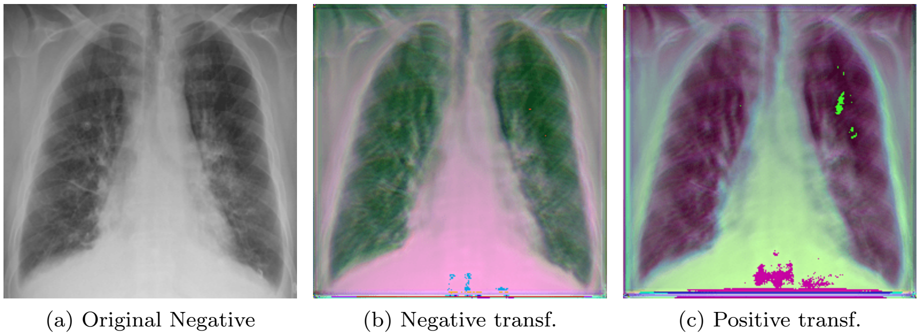


Figure 3: The segmentation process applied in this work.

2. *Class-inherent transformations Network*

   To increase the discrimination capacity of the classification model, we used a Class-inherent transformations (CiT) Network inspired by GANs (Generative Adversarial Networks). This transformation method is actually an array of two generators $G_{\mathbf{P}}$ and $G_{\mathbf{N}}$. $G_{\mathbf{P}}$ learns the inherent-class transformations of the positive class $\mathbf{P}$ and $G_{\mathbf{N}}$ learns the inherent-class transformations of the negative class $\mathbf{N}$. In other words, $G_{\mathbf{P}}$ learns the transformations that bring an input image from its own k domain, with $k \in \{\mathbf{P}, \mathbf{N}\}$, to the $\mathbf{P}$ class domain. While $G_{\mathbf{N}}$ learns the transformations that bring the input image from its k space, with $k \in \{\mathbf{P}, \mathbf{N}\}$, to the $\mathbf{N}$ class space. The classification loss is introduced in the generators to drive the learning of each specific k-class transformations. More details about these transformation networks can be found in [11].
   The architecture of the generators consists of 5 identical residual blocks. Each block has two convolutional layers with 3×3 kernels and 64 feature maps followed by batch-normalization layers and Parametric ReLU

as activation function. The last residual block is followed by a final convolutional layer which reduces the output image channels to 3 to match the inputs dimensions. The classifier is a ResNet-18 which consists of an initial convolutional layer with 7×7 kernels and 64 feature maps followed by a 3×3 max pool layer. Then, 4 blocks of two convolutional layers with 3×3 kernels with 64, 128, 256 and 512 feature maps respectively followed by a 7×7 average pooling and one fully connected layer which outputs a vector of N elements. ReLU is used as activation function.


(a) Original Negative

(b) Negative transf.

(c) Positive transf.

Figure 4: Class-inherent transformations applied to a negative sample. a) Original negative sample; b) Negative transformation; c) Positive transformation

Once the generators learn the corresponding transformations, the dataset is processed using $G_{\mathbf{N}}$ and $G_{\mathbf{P}}$. Two pair of images $(I+, I-)$ will be obtained from each input image I, where $I+$ and $I-$ are respectively the positively and negatively transformed images of I. If I belongs to class $\mathbf{P}$, $G_{\mathbf{P}}$ and $G_{\mathbf{N}}$ will produce the positive transformation $I+ \in \mathbf{P+}$ and the negative transformation $I- \in \mathbf{P\text{-}}$. If an input image I belongs to class $\mathbf{N}$, $G_{\mathbf{P}}$ and $G_{\mathbf{N}}$ will produce its positive $I+ \in \mathbf{N+}$ and negative $I- \in \mathbf{N\text{-}}$ transformations. Figure 4 illustrates with example the transformations applied by $G_{\mathbf{N}}$ and $G_{\mathbf{P}}$. Notice that these transformations are not meant to be interpretable by the human eye but rather help the classification model better distinguish between the different classes.

The original binary problem is then converted into a four classes problem, where the new classes are **N+**, **N-**, **P+** and **P-**.

3. *Learning and inference based on the fusion of CNN twins*

The CNN classification model described above in this section (Resnet-50) is trained to predict the new four classes. The output for each transformed image associated to the original one are actually four tuple. Herein, we propose an inference process to fuse the output. In this way, for each pair $(I+, I-)$, the prediction $\widehat{I}$ of the original image will be either $\mathbf{P}$ or $\mathbf{N}$. Let $\widehat{I+} = \text{argmax } \theta = \text{argmax } (\theta_{\mathbf{N+}}, \theta_{\mathbf{N\text{-}}}, \theta_{\mathbf{P+}}, \theta_{\mathbf{P\text{-}}})$ and $\widehat{I-} = \text{argmax } \psi = \text{argmax } (\psi_{\mathbf{N+}}, \psi_{\mathbf{N\text{-}}}, \psi_{\mathbf{P+}}, \psi_{\mathbf{P\text{-}}})$ be ResNet-50 predictions for $I+$ and $I-$

respectively, where θ and ψ are the probabilities of belonging to each class. Then:

(a) If $\widehat{I+}$ = **N+** and $\widehat{I-}$ = **N-**, then $\widehat{I}$ = **N**.
(b) If $\widehat{I+}$ = **P+** and $\widehat{I-}$ = **P-**, then $\widehat{I}$ = **P**.
(c) If none of the above applies, then

$$\widehat{I} = \begin{cases} \mathbf{N} \text{ if } \max(\theta_{\mathbf{N}_i}, \psi_{\mathbf{N}_i}) > \max(\theta_{\mathbf{P}_i}, \psi_{\mathbf{P}_i}),\ i \in \{+\ ,\ \text{-}\} \\ \mathbf{P} \text{ otherwise.} \end{cases} \quad (1)$$

Experimentally, we used a batch size of 16 and SGD as optimizer.

5. Experiments and Results

In this section we (1) provide all the information about the used experimental setup, (2) evaluate two state-of-the-art COVID classification models on our dataset then, analyze (3) the impact of data pre-processing and (4) Normal-PCR+ severity level on our approach.

5.1. Experimental setup

Due to the high variations between different executions, we performed 5 different 5 fold cross validations in all the experiments. Each experiment uses 80% of COVIDGR 1.0 for training and the remaining 20% for testing. To choose when to stop the training process, we used a random 10% of each training set for validation. In each experiment, a proper set of data-augmentation techniques is carefully selected. All results, in terms of sensitivity, specificity, precision, F1 and accuracy, are presented using the average values and the standard deviation of the 25 executions. The used metrics are calculated as follows:

$$\text{recall(positive class)} = \textit{sensitivity} = \frac{\text{TP}}{\text{actual positives}}$$

$$\text{recall(negative class)} = \textit{specificity} = \frac{\text{TN}}{\text{actual negatives}}$$

$$\text{precision(positive class)} = \frac{\text{TP}}{\text{predicted positives}}$$

$$\text{precision(negative class)} = \frac{\text{TN}}{\text{predicted negatives}}$$

$$\textit{accuracy} = \frac{\text{TP+TN}}{\text{total predictions}}$$

TP and TN refers respectively to the number of true positives and true negatives.

5.2. Analysis of COVIDNet and COVID-CAPS

We compare our approach with the two most related approaches to ours, COVIDNet [17] and COVID-CAPS [19].

- COVIDNet: Currently, the authors of this network provide three versions, namely A, B and C, available at [25]. A has the largest number of trainable parameters, followed by B and C. We performed two evaluations of each network in such a way that the results will be comparable to ours.
  - First, we tested COVIDNet-A, COVIDNet-B and COVIDNet-C, pretrained on COVIDx, directly on our dataset by considering only two classes: Normal (negative), and COVID-19 (positive). The whole dataset (377 positive images and 377 negative images) is evaluated. We report in Table 4 recall and precision results for Normal and COVID-19 classes.
  - Second, we retrained COVIDNet on our dataset. It is important to note that as only a checkpoint of each model is available, we could not remove the last layer of these networks, which has three neurons. We used 5 different 5 fold cross validations. In order to be able to retrain COVIDNet models, we had to add a third Pneumonia class into our dataset. We randomly selected 377 images from the Pneumonia class in COVIDx dataset. We used the same hyper-parameters as the ones indicated in their training script, that is, 10 epochs, a batch size of 8 and a learning rate of 0.0002. We changed covid_weight to 1 and covid_percent to 0.33 since we had the same number of images in all the classes. Similarly, we report in Table 4 recall and precision of our two classes, Normal and COVID-19, and omit recall and precision of Pneumonia class. The accuracy reported in the same table only takes into account the images from our two classes. As with our models, we report here the mean and standard deviation of all metrics.

  Although we analyzed all three A, B and C variations of COVIDNet, for simplicity we only report the results of the best one.

- COVID-CAPS: This is a capsule network-based model proposed in [19] and available at [26]. Its architecture is notably smaller than COVIDNet, which implies a dramatically lower number of trainable parameters. Since the authors also provide a checkpoint with weights trained in the COVIDx dataset, we were able to follow a similar procedure than with COVIDNet:
  - First, we tested the pretrained weights using COVIDx on COVIDGR-1.0 dataset. COVID-CAPS is designed to predict two classes, so we reused the same architecture with the new dataset and compute the evaluation metrics shown in Table 4.
  - Second, COVID-CAPS architecture was retrained over the COVIDGR-1.0 dataset. This process finetunes the weights to improve class separation. The retraining process is performed using the same setup and

hyper-parameters reported by the authors. Adam optimizer is used across 100 epochs with a batch size of 16. Class weights were omitted as with COVIDNet, since this dataset contains balanced classes in training as well as in test. Evaluation metrics are computed for five sets of 5-fold cross-validation test subsets and summarized in Table 4.

Class	Negative		Positive (COVID-19)		Accuracy
Metric	Specificity	Precision	Sensitivity	Precision	
COVIDNet-CXR A [17]	0.27	20	**99.74**	33.78	50
Retrained COVIDNet-CXR A	**89.37±8.88**	60.93±6.20	41.57±17.98	**82.34±8.82**	**65.47±5.53**
COVID-CAPS [19]	26.58	50.78	74.25	50.27	50.41
Retrained COVID-CAPS	64.84± 10.48	**61.76±6.40**	57.89±15.77	62.21±4.86	61.37±5.24

Table 4: COVIDNet and COVID-CAPS results on our dataset

The results from Table 4 show that COVIDNet and COVID-CAPS trained on COVIDx overestimate COVID-19 class in our dataset, i.e., most images are classified as positive, resulting in very high sensitivities but at the cost of low positive predictive value. However, when COVIDNet and COVID-CAPS are re-trained on COVIDGR-1.0 they achieve slightly better overall accuracy and a higher balance between sensitivity and specificity, although they seem to acquire a bias favoring the negative class. In general, none of these models perform adequately for the detection of the disease from CXR images in our dataset.

5.3. Results and Analysis of COVID prediction

The results of the baseline COVID classification model considering all the levels of severity, with and without segmentation; and COVID-SDNet are shown in Table 5.

Class	N			P			Accuracy
Metric	**Specificity**	Precision	F1	**Sensitivity**	Precision	F1	
COVIDNet-CXR	89.37±8.88	60.93±6.20	71.84±2.94	41.57±17.98	82.34±8.82	52.27±14.89	65.47±5.53
COVID-CAPS	64.84± 10.48	61.76±6.40	62.44±4.97	57.89±15.77	62.21±4.86	58.81±10.65	61.37±5.24
Without seg.	75.25±6.78	71.04±3.13	72.84±2.87	68.95±6.27	74.04±4.45	71.09±2.88	72.10±2.31
With seg.	71.37±9.25	73.89±5.41	71.97±4.39	73.68±9.33	72.59±4.39	72.63±4.19	72.54±3.19
COVID-SDNet	**79.20±6.29**	**76.58±3.92**	**77.67±3.21**	**75.43±5.91**	**78.82±5.04**	**76.82±3.08**	**77.31±2.92**

Table 5: Results of COVID prediction using ResNet-50 with and without segmentation, COVID-SDNet, Retrained COVIDNet-CXR A and Retrained COVID-CAPS. All four levels of severity in the positive class are taken into account.

In general, COVID-SDNet achieves better and more stable results than the rest of approaches. In particular, COVID-SDNet achieved the highest balance between specificity and sensitivity with 77.67 ± 3.21 F1 in the negative class and 76.82 ± 3.08 F1 in the positive class. Most importantly, COVID-SDNet achieved the highest specificity with 79.20 ± 6.29, sensitivity 75.43 ± 5.91 and accuracy with 77.31 ± 2.92. When comparing the results of the baseline classification model with and without segmentation, we can observe that the use of segmentation improves substantially the sensitivity which is the most important criteria for a triage system. This can be explained by the fact that segmentation allows the model to focus on most important parts of the CXR image.

Analysis per severity level

To determine which levels are the hardest to distinguish by the best approach, we have analyzed the accuracy per severity level (S), with accuracy(S) = $\frac{\text{Correct predictions(S)}}{\text{Total number(S)}}$, where $S = \{\text{Normal-PCR+, Mild, Moderate, Severe}\}$. The results are shown in Table 6.

S (Severity level)	accuracy (S)(%)
Normal-PCR+	38.68 ± 2.44
Mild	66.5 ± 8.04
Moderate	88.14 ± 2.02
Severe	97.37 ± 1.86

Table 6: Results of COVID-SDNet per severity level.

As it can be seen from these results, COVID-SDNet correctly distinguish Moderate and Severe levels with an accuracy of $88,14\%$ and $97,37\%$ respectively. This is due to the fact that Moderate and Severe CRX images contain more important visual features than Mild and Normal-PCR+ which ease the classification task. Normal-PCR+ and Mild cases are much more difficult to identify as they contain few or none visual features. These results are coherent with the clinical studies provided in [6] and [3] which report that expert sensitivity is very low in Normal-PCR+ and Mild infection levels. Recall that the expert eye does not see any visual signs in Normal-PCR+ although the PCR is positive. Those cases are actually considered as asymptomatic patients.

5.4. Analysis of the impact of Normal-PCR+

To analyze the impact of Normal-PCR+ class on COVID-19 classification, we trained and evaluated the baseline model, COVID-SDNet classification stage, COVIDNet-CXR-A and COVID-CAPS, on COVIDGR by eliminating Normal-PCR+. The results are summarized in Table 7.

Class	N			P			Accuracy
Metric	**Specificity**	Precision	F1	**Sensitivity**	Precision	F1	
COVIDNet-CXR	90.14±9.73	63.24±7.71	73.50±3.97	50.51±18.31	78.75±12.81	59.25±14.70	70.32±5.96
COVID-CAPS	72.16±7.04	66.01±5.94	68.64±4.42	61.91±10.97	69.16±5.29	64.81±7.44	67.04±5.03
With seg.	80.28±6.98	77.12±4.93	78.33±3.36	75.47±8.11	79.78±4.87	77.16±4.16	77.87±3.29
COVID-SDNet	**81.06±5.32**	**81.58±4.76**	**81.15±3.34**	**81.33±5.94**	**81.34±4.17**	**81.16±3.56**	**81.20±3.32**

Table 7: Results of the baseline classification model with segmentation, COVID-SDNet, retrained COVIDNet-CXR-A and retrained COVID-CAPS. Only three levels of severity are considered, Mild, Moderate and Severe.

Overall, all the approaches systematically provide better results when eliminating Normal-PCR+ from the training and test processes, including COVIDNet-CXR-A and COVID-CAPS. In particular, COVID-SDNet still represents the best and most stable approach.

Analysis per severity level

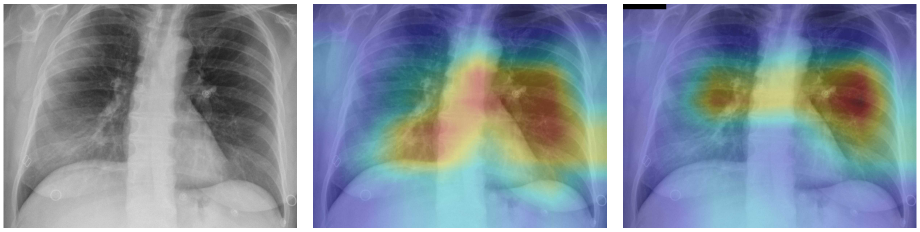
A further analysis of the accuracy at the level of each severity degree (see Table 8) demonstrates that eliminating Normal-PCR+ decreases the accuracy in Mild and Moderate severity levels by 10% and 3.75% respectively.

S (Severity level)	accuracy (S)(%)
Mild	59.5 ± 3.22
Moderate	84.83 ± 2.51
Severe	97.63 ± 0.98

Table 8: Results of COVID-SDNet by severity level without considering Normal-PCR+.

These results show that although Normal-PCR+ is the hardest level to predict, its presence improves the accuracy of lower severity levels, especially Mild level.

6. Inspection of model's decision

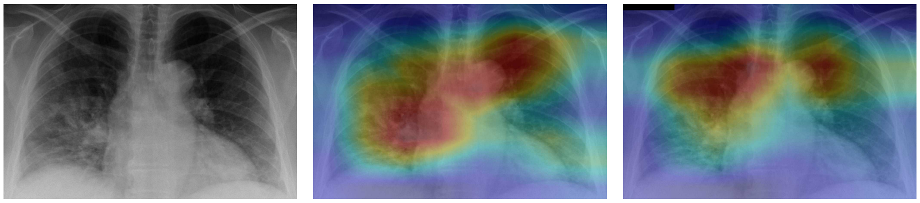


(a) Original Positive (Mild)

(b) why positive

(c) why negative

Figure 5: Heatmap showing the parts of the input image that triggered the positive prediction (b) and counterfactual explanation (c)

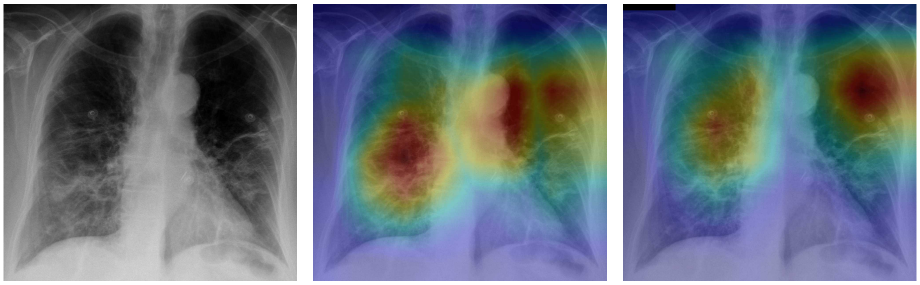


(a) Original Positive (Moderate)

(b) why positive

(c) why negative

Figure 6: Heatmap showing the parts of the input image that triggered the positive prediction (b) and counterfactual explanation (c)

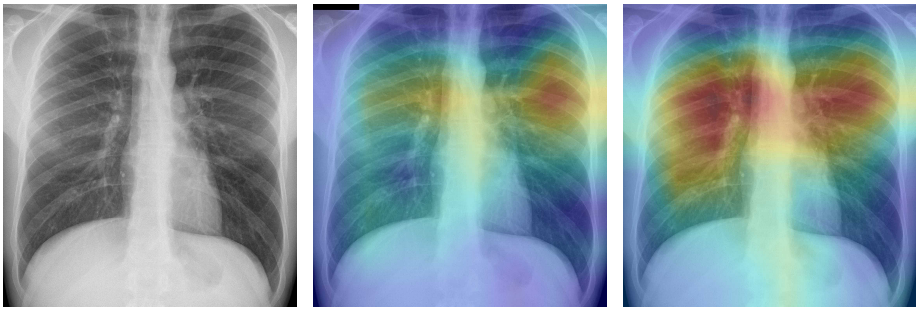


(a) Original Positive (Severe)

(b) why positive

(c) why negative

Figure 7: Heatmap showing the parts of the input image that triggered the positive prediction (b) and counterfactual explanation (c)



(a) Original Negative

(b) why positive

(c) why negative

Figure 8: Heatmap that explains the parts of the input image that triggered the counterfactual explanation (b) and the negative actual prediction (c).

Automatic DL diagnosis systems alone are not mature yet to replace expert radiologists. To help clinician making decisions, these tools must be interpretable so that clinicians can decide whether to trust the model or not [27]. We inspect what led our model make a decision by showing the regions of the input image that triggered that decision along with its counterfactual explanation by showing the parts that explain the opposite class. We adapted Grad-CAM method [28] to explain the decision of the negative and positive class.

Figures 5, 6 and 7 show (a) the original CXR image, (b) visual explanation by means of a heat-map that highlights the regions/pixels which led the model to output the actual prediction and (c) its counterfactual explanation using a heat-map that highlights the regions/pixels which had the highest impact on predicting the opposite class. The larger high intensity areas in the heat-map determine the final class. However, Figure 8(b) represents first the counterfactual explanation and Figure 8(c) represents the explanation of the actual decision.

As expected, negative and positive interpretations are complementary, i.e, areas which triggered the correct decision are opposite, in most cases, to the areas that triggered the decision towards negative. In CXR images with different severity levels, the heat-maps correctly point out opaque regions due to different levels of infiltrates, consolidations and also to osteoarthritis.

In particular, in Figure 5(b), the red areas in the right lung points out a region with infiltrates and also osteoarthritis in the spine region. Figure 6 (b) correctly shows moderate infiltrates in the right lower and lower-middle lung fields in addition to a dilation of ascending aorta and aortic arch (red color in the center). Figure 5(c) shows normal upper-middle fields of both lungs (less important on the left due to aortic dilation). Figure 7(b) indicates an important bilateral pulmonary involvement with consolidations.

As it can be observed in Figure 8(c), the explanation of the negative class correctly highlights a symmetric bilateral pattern that occupies a larger lung volume especially in regions with high density. In fact, a very similar pattern is shown in the counterfactual explanation of the positive class in Figures 5(c), 6(c) and 7(c).

7. Conclusions

This paper introduced a dataset, named COVIDGR, with high clinical value. COVIDGR includes the four main COVID severity levels identified by a recent radiological study [3]. We proposed a methodology, called COVID-SDNet, that combines segmentation, data-augmentation and data transformation. The obtained results show the high generalization capacity of COVID-SDNet, especially on severe and moderate levels as they include important visual features. The existence of few or none visual features in Mild and Normal-PCR+ reduces the opportunities for improvement.

As main conclusions, we must highlight that COVID-SDNet can be used in a triage system to detect especially moderate and severe patients. Finally, we must also mention that more robust and accurate triage system can be built by fusing our approach with other approaches such as the one proposed in [29].

As future work, we are working on enriching COVIDGR with more CXR images coming from different hospitals. We are planning to explore the use of additional clinical information along with CXR images to improve the prediction performance.

Acknowlegments

This work was supported by the project DeepSCOP-Ayudas Fundacin BBVA a Equipos de Investigacin Cientfica en Big Data 2018 and the Spanish Ministry of Science and Technology under the project TIN2017-89517-P. S. Tabik was supported by the Ramon y Cajal Programme (RYC-2015-18136). A. Gmez-Ros was supported by the FPU Programme 998758-2016. E.G was supported by the European Research Council (ERC Grant agreement 647038 [BIODESERT])

Ethics

This project is approved by the Provincial Research Ethics Committee of Granada.

References

[1] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, Projecting the transmission dynamics of sars-cov-2 through the postpandemic period, Science.

[2] American college of radiology and others. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected covid-19 infection, https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection (2020).

[3] 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. Chung, et al., Frequency and distribution of chest radiographic findings in covid-19 positive patients, Radiology (2020) 201160.

[4] Y. Li, L. Yao, J. Li, L. Chen, Y. Song, Z. Cai, C. Yang, Stability issues of rt-pcr testing of sars-cov-2 for hospitalized patients clinically diagnosed with covid-19, Journal of Medical Virology.

[5] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest ct for covid-19: comparison to rt-pcr, Radiology (2020) 200432.

[6] M. B. Weinstock, A. Echenique, J. W. R. DABR, A. Leib, F. A. ILLUZZI, Chest x-ray findings in 636 ambulatory patients with covid-19 presenting to an urgent care center: A normal chest x-ray is no guarantee, J Urgent Care Med,(14 (7)) (2020) 13–18.

[7] 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.

[8] J. P. Cohen, P. Morrison, L. Dao, Covid-19 image data collection, arXiv 2003.11597.
URL https://github.com/ieee8023/covid-chestxray-dataset

[9] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, F. Herrera, Big Data Preprocessing - Enabling Smart Data, Springer, 2020. doi:10.1007/978-3-030-39105-8.

[10] S. Tabik, D. Peralta, A. Herrera-Poyatos, F. Herrera, A snapshot of image pre-processing for convolutional neural networks: case study of mnist, International Journal of Computational Intelligence Systems 10 (1) (2017) 555–568.

[11] M. Rey-Area, E. Guirado, S. Tabik, S. Ruiz-Hidalgo, Fucitnet: Improving the generalization of deep learning networks by the fusion of learned class-inherent transformations, arXiv preprint arXiv:2005.08235.

[12] Radiological society of north america. rsna pneumonia detection challenge. URL `https://www.kaggle.com/c/rsnapneumonia-detection-challenge/data`

[13] C. et al., Figure 1 covid-19 chest x-ray dataset initiative. URL `https://github.com/agchung/Figure1-COVID-chestxray-dataset`

[14] 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.

[15] A. E. Johnson, T. J. Pollard, S. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, Mimic-cxr: A large publicly available database of labeled chest radiographs, arXiv preprint arXiv:1901.07042.

[16] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, arXiv preprint arXiv:1901.07441.

[17] Z. Q. L. Linda Wang, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images (2020). `arXiv:2003.09871`.

[18] S. Kundu, H. Elhalawani, J. W. Gichoya, C. E. Kahn Jr, How might ai and chest imaging help unravel covid-19s mysteries? (2020).

[19] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, A. Mohammadi, Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images, arXiv preprint arXiv:2004.02696.

[20] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, U. R. Acharya, Automated detection of covid-19 cases using deep neural networks with x-ray images, Computers in Biology and Medicine (2020) 103792.

[21] M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan, et al., Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images, arXiv preprint arXiv:2004.04582.

[22] B. Ghoshal, A. Tucker, Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection, arXiv preprint arXiv:2003.10769.

[23] I. D. Apostolopoulos, T. A. Mpesiana, Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Physical and Engineering Sciences in Medicine (2020) 1.

[24] U-Net lung segmentation, Accesible en: `https://www.kaggle.com/eduardomineo/u-net-lung-segmentation-montgomery-shenzhen` (2020).

[25] COVIDNet, Accesible en: `https://github.com/lindawangg/COVID-Net` (2020).

[26] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, A. Mohammadi (2020). [link].
URL `https://github.com/ShahinSHH/COVID-CAPS`

[27] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[29] J. P. Cohen, L. Dao, P. Morrison, K. Roth, Y. Bengio, B. Shen, A. Abbasi, M. Hoshmand-Kochi, M. Ghassemi, H. Li, T. Q. Duong, Predicting covid-19 pneumonia severity on chest x-ray with deep learning, arXiv preprint arXiv:2005.11856.