IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 10, 2021, accepted March 7, 2021, date of publication March 10, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3065456

Bias Analysis on Public X-Ray Image Datasets of Pneumonia and COVID-19 Patients

OMAR DEL TEJO CATALÁ[1], ISMAEL SALVADOR IGUAL[1], FRANCISCO JAVIER PÉREZ-BENITO[1], DAVID MILLÁN ESCRIVÁ[1], VICENT ORTIZ CASTELLÓ[1], RAFAEL LLOBET[1,2], AND JUAN-CARLOS PERÉZ-CORTÉS[1,3]

[1]Instituto Tecnológico de Informática (ITI), Universitat Politècnica de València, 46022 Valencia, Spain
[2]Department of Computer Systems and Computation (DSIC), Universitat Politècnica de València, 46022 Valencia, Spain
[3]Department of Computing Engineering (DISCA), Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Ismael Salvador Igual (issalig@iti.upv.es)

This work was supported by Generalitat Valenciana through the "Instituto Valenciano de Competitividad Empresarial—IVACE" under Grant IMDEEA/2020/69.

ABSTRACT Chest X-ray images are useful for early COVID-19 diagnosis with the advantage that X-ray devices are already available in health centers and images are obtained immediately. Some datasets containing X-ray images with cases (pneumonia or COVID-19) and controls have been made available to develop machine-learning-based methods to aid in diagnosing the disease. However, these datasets are mainly composed of different sources coming from pre-COVID-19 datasets and COVID-19 datasets. Particularly, we have detected a significant bias in some of the released datasets used to train and test diagnostic systems, which might imply that the results published are optimistic and may overestimate the actual predictive capacity of the techniques proposed. In this article, we analyze the existing bias in some commonly used datasets and propose a series of preliminary steps to carry out before the classic machine learning pipeline in order to detect possible biases, to avoid them if possible and to report results that are more representative of the actual predictive power of the methods under analysis.

INDEX TERMS
Deep learning, COVID-19, convolutional neural networks, chest X-ray, bias, segmentation, saliency map.

I. INTRODUCTION

Chest X-ray (CXR) radiography is the most widely accepted imaging modality for detecting pneumonia and it is becoming crucial for tracking the clinical evolution of COVID-19 patients [1]. The COVID-19 disease is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and has become a global pandemic in a few months. Early diagnosis is a key factor due to the stealthy contagious nature of the virus and a lack of vaccines or effective treatments and, thus, it helps to prevent further spreading and to control it under the existing healthcare facilities. The small size of the acquisition devices, their ease of operation and their low cost make them more widely available than the Computer Tomography (CT) equipment, despite image quality and the diagnostic performance of CT are superior.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

As a response to the COVID-19 outbreak, the scientific community has rapidly reacted and a lot of works using CXR images for COVID-19 detection have been published. The majority of them make use of well-known CNN architectures such as VGG [2], ResNet [3]–[5], SqueezeNet [3], [6], DenseNet [7] and also combine them with decision trees [8] and Support Vector Machines (SVM) [9]. Given the difficulty of obtaining COVID-19 samples, GAN networks have been used [10], [11] in order to enhance the performance. Moreover, other approaches [12], [13] based on multi-resolution methods report results that are comparable to those obtained by CNNs.

Machine learning models need large amounts of data which, in this case, are difficult to acquire, being the existing collections a mix of already well-known datasets and new COVID-19 image datasets. This heterogeneous mixture of observations provides more variety and usually reduces epistemic uncertainty. However, if these datasets, for instance, are

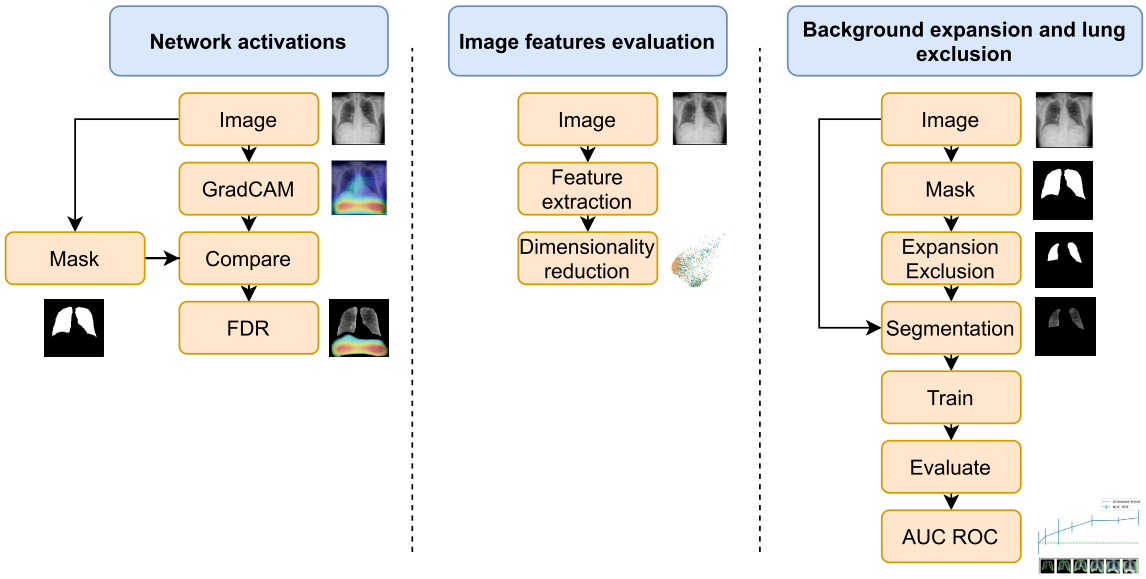


FIGURE 1. Workflow of the different experiments. From left to right: Network activations, Image features evaluation and background expansion and lung exclusion.

not equally balanced (label-wise), they may induce a certain amount of dataset bias to the training phase. This happens when the images can be easily discriminated by features not relevant to the task, i.e. if the dataset inadvertently contains some distinctive features which are not related to the disease and are not shared among the source datasets. For instance, let's assume an extreme case. Two image datasets are formed by two different classes, that is, dataset A made of class A samples and dataset B of class B samples. Let's assume in most dataset A samples there is a white rectangle on the top right corner, and the true class features are not as trivial. Classifiers will focus on the easiest feature to discriminate between classes and not the true class features. Therefore, this leads to poor generalization; given a new dataset C full of class A, samples with no white rectangle will be misclassified.

We have detected significant biases in some of the most commonly used datasets intended for pneumonia and COVID-19 detection and we suspect that the accuracy reported in some studies might be due in part to them, and thus not directly related to the image features that could characterize the disease. These biases could arise, for example, when using some specific devices to acquire images of patients with a low probability of suffering the disease (mainly controls), and different ones for those patients with a high probability of suffering it (mainly cases). This could happen, for example, when most of the patients are screened in certain health services and highly suspicious patients are derived to a different area or, even worse, when, aiming to increase the number of controls or cases, a dataset is expanded with samples coming from significantly different origins and labeled with unbalanced class identifiers. In these cases, a CNN trained to discriminate between cases and controls could learn to differentiate images from different origins rather than finding features actually related to the disease.

Therefore, to effectively assess the performance of the classifier, there must exist a previous study of the dataset bias, so that the results can be validated. Thus, we present several studies to assess the validity of the results. The following datasets will be used to perform the experiments: BIMCV Padchest, CheXpert, RSNA and a COVID-19 image data collection that we will refer to as COVIDcxr, which will be further described in Section II-A.

The main contributions of this work are:

- To propose a bias analysis methodology to assert the validity of the results achieved on a dataset.
- To study the possible existence of bias in three broadly used pneumonia classification datasets.
- To study the effect of mixing several datasets.

This work is structured as follows: Section I outlines the problem of bias in CXR datasets. After that, the datasets and networks used, along with the proposed methodology are described in Section II. The workflow related to this section can be seen in Figure 1. Section III shows the results achieved using this article's methodology over the proposed datasets and Section IV gives an analysis of the results. Finally, conclusions are presented in Section VI.

II. METHODS

A. DATASETS

Several public datasets have been used in this article:

- PADCHEST[1] [14] is a CXR dataset that includes more than 160K images from 67625 patients that were reported by radiologists at Hospital de San Juan (Spain)

[1] http://bimcv.cipf.es/bimcv-projects/padchest/

from 2009 to 2017. The reports are labeled with 174 different radiographic findings, 19 differential diagnoses and 104 anatomic locations. 27% of the reports were manually annotated by trained physicians and the remaining set was labeled using a supervised method based on a Recurrent Neural Network with attention mechanisms. Generated labels were validated, achieving a 0.93 Micro-F1 score using an independent test set. For the experiments, only Posterior-Anterior images are considered. Therefore, there are 9110 images in the remaining dataset: 6790 control and 2320 pneumonia images.

- RSNA pneumonia dataset[2] is made up of images from the National Institutes of Health (NIH) and labeled by the Radiological Society of North America along with the Society for Thoracic Radiology and MD.ai. The goal of this dataset was to develop an AI classifier capable of distinguishing between pneumonia and control images, so it was released in a Kaggle competition in 2018. It consists of 26684 images from which 20672 are control and 6012 are pneumonia images.
- CheXpert dataset[3] [15] is provided by Stanford University and contains 224316 chest radiographs of 65240 patients with labels of 14 sub-categories. The exams were performed at Stanford Hospital between October 2002 and July 2017. Structured labels for the images were created by an automated rule-based labeler, which the researchers developed to extract observations from free-text radiology reports. From the 224316 chest radiographs, this article only takes the ones related to pneumonia and control cases. Therefore, 5870 images are remaining in the dataset: 4878 control and 992 pneumonia images.
- COVID-19 image data collection (COVIDcxr)[4] [16] is a project to collect X-ray and CT images that present COVID-19, SARS, MERS and ARDS from online sources. These sources are varied: scientific publications, websites, etc. As of June 2020, COVIDcxr has around 424 COVID-19 images and is one of the largest COVID-19 datasets publicly available to the best of our knowledge.

B. MOTIVATION

The motivation for this study comes from analyzing the results of a neural network trained to classify between radiographic images of patients with pneumonia and healthy control patients in order to determine the validity of the classification. An interesting first validation can be done by visualizing the network's activation heatmaps. When we performed these checks against networks trained with pneumonia datasets, we observed many suspicious patterns, as these heatmaps often highlighted areas of the image which did not contain lung tissue (see Figure 2). This made us suspect that the networks were learning to classify, achieving large values of AUC ROC, using features unrelated to the task. Thus, the datasets might be biased.

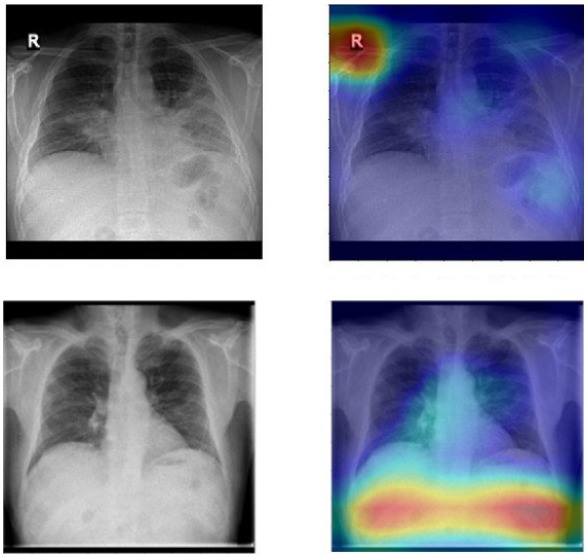


FIGURE 2. Lung heatmaps for BIMCV's dataset.

Grad-CAM [17] allows us to visualize the gradient of the label in the final convolutional layer to produce a heatmap depicting regions of the image that are relevant for the prediction. Blue pixels and red pixels correspond to low and high values of the gradient at the final convolutional layer, respectively.

As observed in Figure 2, there are highly activated regions in areas without lung presence when the expected activation should be inside the lung. It is not known how many pixels inside the lungs should show an activation, as no detection mask is available. However, we can assume that the activation map in a control patient should not exceed a given threshold, whilst a positive case's map should show widespread activations within the lungs. Nonetheless, the activated area outside the lungs should be minimal in all cases. For this reason, a measure to inform about the distribution of the activated pixels could be useful.

Given a heatmap image $I = \{p_{ij}\} \in \mathrm{Mat}_{n,m}(\mathbb{R})$, where n is the number of rows, m the number of columns, and p_{ij} represents the pixel value at row i and column j. Let A be a region of interest and B its complement. Let t be the activation map threshold, and let R and W be the number of pixels with an activation value higher than t that are in A and B respectively.

We can calculate the percentage of pixels with an activation value over a threshold that fall outside an expected region as the quotient between W and $W + R$ (see Figure 3 and the equations below, where $p \in \{p_{ij}\} = I$).

Considering activated pixels in region W as false positives (FP) and activated pixels in region R as true positives (TP), the above quotient corresponds to the False Discovery

[2]https://www.kaggle.com/c/rsna-pneumonia-detection-challenge
[3]https://www.healthimaging.com/topics/artificial-intelligence/stanford-researchers-release-chest-x-ray-dataset-train-ai
[4]https://github.com/ieee8023/covid-chestxray-dataset

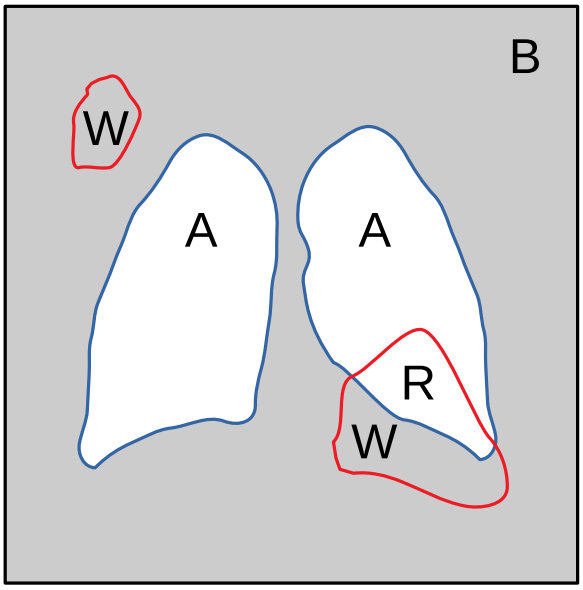


FIGURE 3. Activation regions diagram.

Rate (FDR), which is the complement of the Positive Predictive Value (PPV).

$$R = \text{TP} = |(p > t) \cap A|$$
$$W = \text{FP} = |(p > t) \cap B|$$
$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}}$$
$$\text{FDR} = 1 - \text{PPV}$$

For instance, in this task, any activated pixel that falls outside the lungs is marked as wrong (W), as no information should be found there. The lower this value, the better. This score is designed to measure the validity of the trained CNN classifier based on its activation maps and allows the selection of different operation points depending on the threshold t to be applied to the heatmaps. In this work, t is set to 90% of the maximum heatmap value.

Table 1 shows the computed FDR for the activation maps under three different datasets. It is worth noting that some image findings are usually located on the border of the lungs, so if the highlighted area is near the border, some pixels might easily fall outside the region (A) and be considered as wrong (W). On the grounds of the information provided by the FDR, further experiments would be required to measure the extent to which this phenomenon affects the datasets.

TABLE 1. False discovery rate of activation maps for three different datasets.

	FDR
BIMCV	71.83%
RSNA	40.46%
CheXpert	50.68%

Additionally, some suspicious patterns appeared when visualizing the grayscale histograms of the images. Ideally, gray levels of images from different sources should be equally distributed, but in practice, this may not happen and give rise to inaccurate conclusions. The histograms of the images may be considered as Probability Density Functions (PDFs) and may serve to measure the variability among gray-level distributions using a methodology based on information geometry [18]. This methodology has been successfully applied to characterize EHR (Electronic Health Record) data [19], [20], to assess the variability among patients with different headache pain intensity [21], or to detect pixel distribution differences among images acquired from different mammographs [22].

Given a set of PDFs, this approach is based on the computation of the distance between each pair of PDFs using the Jensen-Shannon distance. The simplex where each point represents a PDF and the distance between two points is the Jensen-Shannon distance between the two PDFs they represent is known as a statistical manifold, which in turn is a Riemannian manifold. For visualization purposes, this simplex may be embedded in a real Euclidean space by using Multidimensional Scaling [23] and, finally, projected into two dimensions using a dimension reduction algorithm such as Principal Component Analysis.

This methodology was applied three times to a random balanced sample of 2000 individuals (1000 pneumonia cases and 1000 controls) of each dataset mentioned, which will be described in section II-A. Firstly, it was applied to the histograms of the complete images and, after a segmentation step, which will be described in detail in section II-D, the variability analysis was applied only to the histograms of the backgrounds, and then to the histograms of the lungs (see Figure 4). The variability of the three datasets is shown in Figure 5.

In the center row of Figure 5, which depicts the distributions of the backgrounds of the different datasets, we can see that the first two columns show distinct clusters composed predominantly of cases or controls that allow a certain degree of discrimination without taking into account the lung tissue. In fact, the last row, which represents lung area, shows fewer differences between the cases and control patient histograms. In the last column, corresponding to CheXpert's dataset, these differences are not evident.

This could imply that, for some datasets, as BIMCV and RSNA, a Machine Learning algorithm can classify pneumonia and control cases using features outside the lungs.

C. NETWORK

In this article, Convolutional Neural Networks (CNNs) are used to classify the CXR images. These Machine Learning models have been widely employed in the last years for image classification, particularly in the field of medical imaging. The CNN topology used is VGG16 [24], which is broadly reported as a good classifier for chest image analysis [25]–[27]. In this scenario, a common practice with this type of networks is to trim the last layers (usually dense layers) and add a lighter classifier, which in this

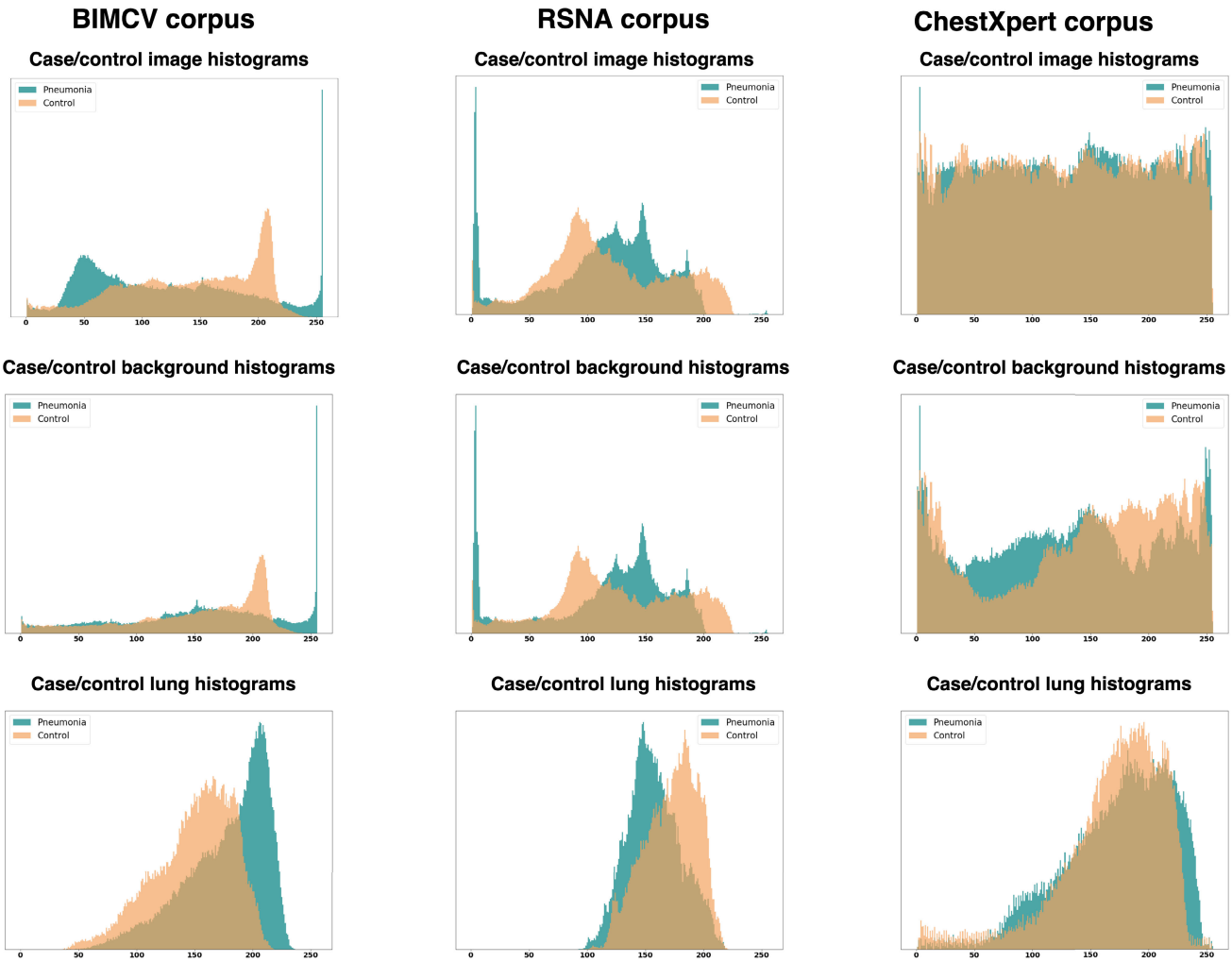


FIGURE 4. Example of case and control patient histograms. The first row shows the histogram of the whole image for an example of a case and a control patient, the second row shows the histogram of the background (the image with the lung area subtracted) and, the last one shows the histogram of the lungs.

case is a Global Average Pooling followed by a Multilayer Perceptron, which projects the pooled features of VGG's last convolution to 64 dimensions before performing the classification.

Transfer learning technique is a common practice within Deep Learning models. It is proven that pretrained networks, in particular their first layers, are generic and can be transferred to new domains without requiring special training. In fact, it also facilitates training for domains with a scarce amount of training samples. Therefore, the VGG16 network used is pretrained with Imagenet dataset, and the last 2 convolutional layers, along with the classification layers, are unfrozen for domain training.

It is noteworthy that the network structure is, up to a point, not critical for the conclusions drawn in this article, as it is not trying to present advancement in the state-of-the-art classification for the datasets used. The focus is rather on comparing the results obtained for images coming from different datasets, and whether those results suggest the presence of classification biases within the data. Nonetheless, it must at least achieve an acceptable accuracy in order to ensure the extracted features are good enough and close to the ones extracted in other articles.

D. SEGMENTATION

By segmenting the lungs, it is possible to remove parts of the image that do not contain relevant information and that can be a source of noise or bias, such as the presence of text annotations that can identify a machine or a hospital, or the appearance of images coming from specific medical devices that have been used in more cases than control patients or vice versa.

Lung segmentation in CXR images has been successfully tackled with different approaches during the last years [28]. For this work, a U-Net network has been trained on the Montgomery dataset [29]. Moreover, we have manually labeled a total of 1115 images coming from BIMCV's

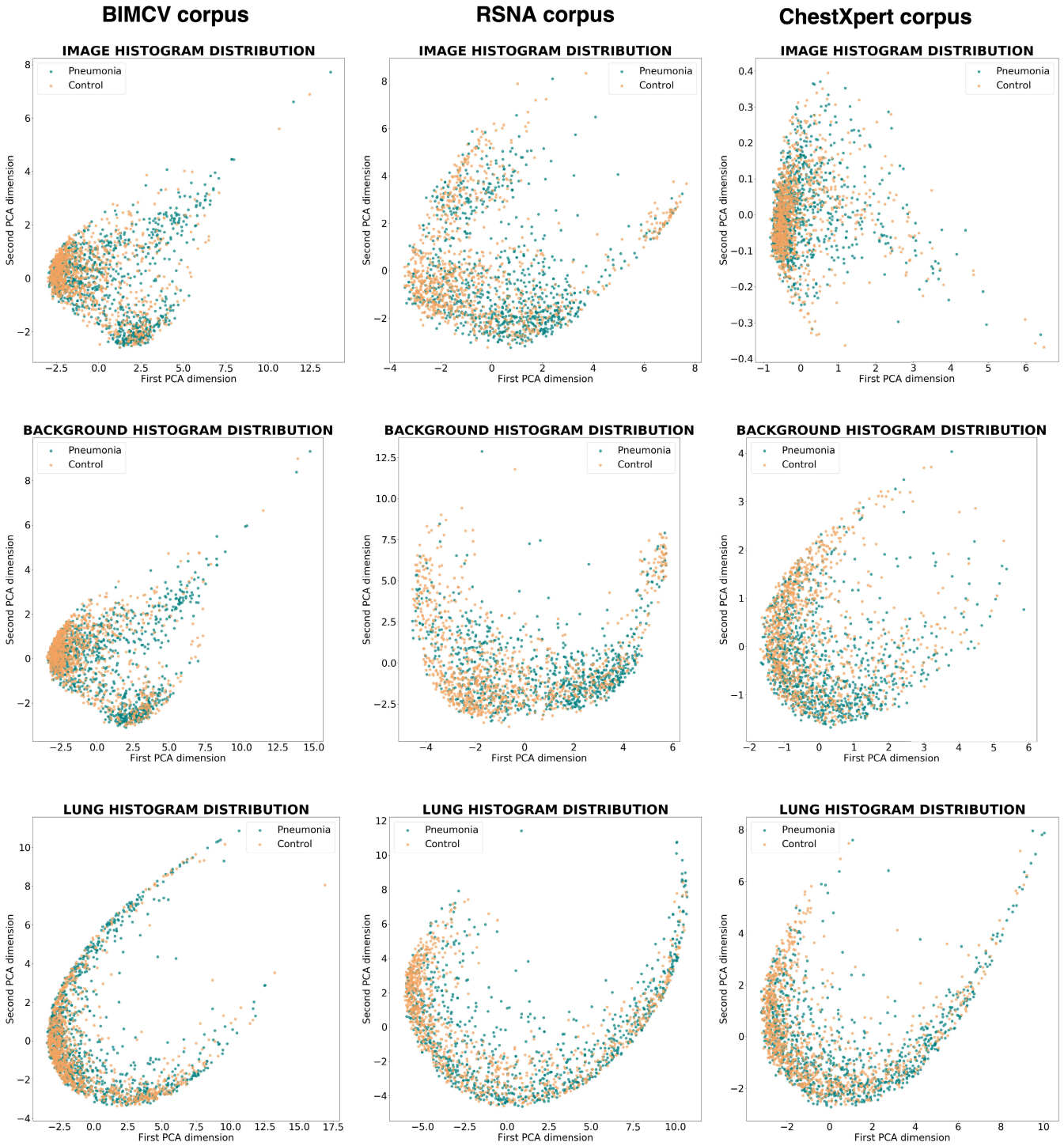


FIGURE 5. **Image histogram variability. The first row represents the variability of the histograms of the complete images, the second row the variability of the background histograms (the images with the lung area subtracted), and the third row the histograms of the lungs. The first column represents a sample of BIMCV's dataset, the second column a sample of RSNA's, and the last, a sample of CheXpert's.**

Padchest dataset to increase the number of training images. Figure 6 shows the segmentation results. This network achieves 0.974 DICE and 0.934 IoU scores over the Montgomery test partition, where DICE and IoU are defined as follows, being A and B the predicted segmentation mask and the true segmentation mask.

$$\text{DICE} = \frac{2\,|\,A \cap B\,|}{|\,A\,| + |\,B\,|}$$

$$\text{IoU} = \frac{A \cap B}{A \cup B}$$

E. BIAS ANALYSIS

This work proposes a methodology to measure the degree of bias in a dataset. The focus is on the classification of pneumonia or COVID against control samples, but the methods can be generalized to other classification tasks where prior knowledge of the region of interest is available.

As stated before, areas that should not contain information about the problem can be possibly used to discriminate between classes, for example, text annotations or image features related to the medical devices employed. In order to solve this problem, we make use of a segmentation algorithm to extract the relevant regions which in this case are the lungs (see Figure 6). These regions will be referred to as masks. The rest of the image will be considered as background (see *B* in Figure 3).

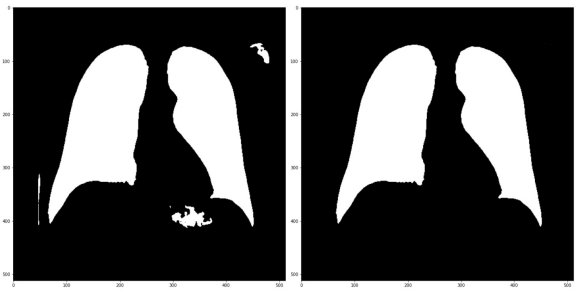


FIGURE 6. Lung segmentation (left) and after post-process (right).

To check the previous hypothesis presented in II-B, two experiments were carried out by training a model with different image areas according to the following ideas:

- We want to study how the background affects the results. Starting from an image that contains only the lungs (the background is erased), the visible region is progressively expanded to include more background by means of sequential dilation operations over the mask (see Figure 7a). An unbiased dataset should not increase the classification accuracy along this process.
- We want to analyze how the lack of lung area affects the results; this time starting from the whole image and progressively removing the lungs (see Figure 7b). The classification accuracy over an unbiased dataset should progressively drop from its maximum value (whole image) to 0.5 AUC ROC.

Thus, adjusting the expansion or exclusion of the lung region will allow us to trace the variation of the accuracy metric. We used images scaled to 256 × 256 pixels. For background expansion, lung segmentation masks were dilated 0, 10, 30, 50, 80, 120 and 140 pixels and for lung exclusion, masks were eroded 0, 10, 20, 30, 40 and 100 pixels (from right to left in Figure 7).

Figure 7a shows the lung segmented area in blue and the background expansion in green. Also, Figure 7b shows the lung exclusion area in yellow. Additionally, a detailed workflow for this experiment is shown in Figure 8

F. COMBINATION ANALYSIS

Combining datasets can be useful to enlarge the sample size, increase the variability explained by the data, and reduce the epistemic uncertainty of the classifiers. This latter is related to the problem-domain knowledge of the model, being it the uncertainty or lack of knowledge bound to the limited amount of data. However, if the combination and the balance among the classes are not carefully controlled, a classifier may learn to discriminate between features of the different datasets.

To check this hypothesis, we mixed RSNA and CheXpert datasets to achieve a balanced combination by adding positive pneumonia observations from the RSNA dataset into CheXpert. The latter is a highly unbalanced dataset (83% of negative and 27% of positive observations after our pre-process and segmentation validity filters), so it could be considered a good idea to add positive samples from another dataset. Needless to say, if the images from RSNA have distinct features that allow the classifier to tell them apart from CheXpert, for example including a large proportion of images from a particular equipment brand or model, the system will learn to classify the images from that equipment as positive, regardless of any image content that could be related to the disease.

Additionally, we simulated the combination of COVID-19 and control datasets and evaluated their bias with the proposed method. In particular, the datasets combined are positive COVID-19 cases from COVIDcxr with CheXpert's negative control samples. COVIDcxr is built with datasets from different origins, hence this experiment illustrates the likely problematic effects of heterogeneous data combinations.

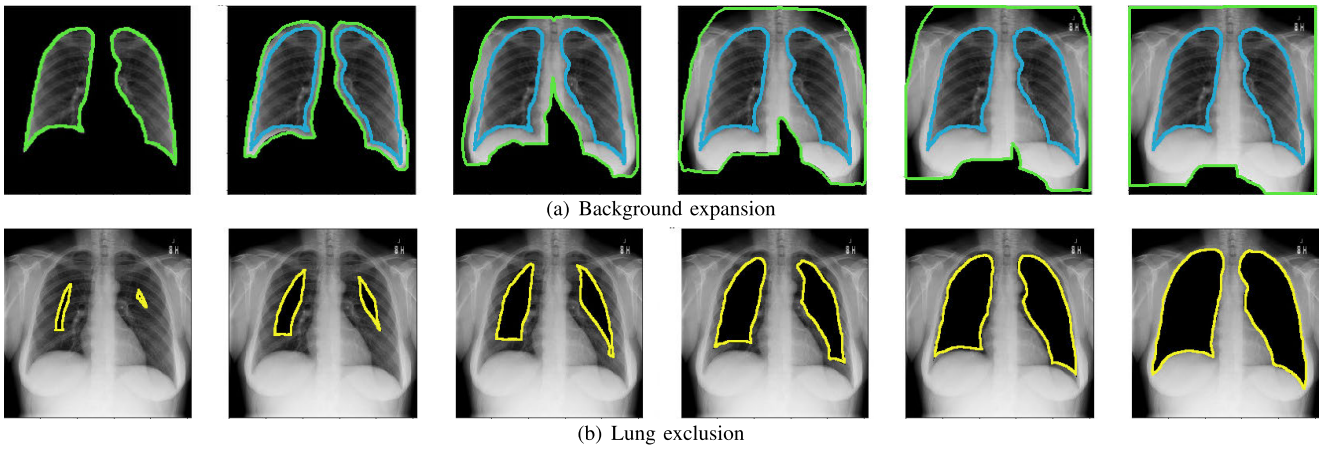
Based on our methodology that probes the discrimination induced outside the lungs, the expectations about the results of the experiment, if there is bias in the dataset, are: (1) the background expansion could increase the accuracy and (2) the accuracy when occluding the lungs should differ significantly from the 0.5 AUC ROC. Did the results follow these predictions, the hypothesis would be confirmed.

III. RESULTS

A. BACKGROUND EXPANSION AND LUNG EXCLUSION STUDY

In the previous section, we proposed to examine the performance of classification experiments varying the addition of background and the reduction of the lung area. The expected results of the first test for a non-biased dataset, where the background area is added to the initial lung-only images, is that the classification rate stays constant (or almost constant, due to possible imprecise segmentation and other random perturbations), as the disease information is already present from the beginning.

In the second scenario, the accuracy should potentially drop from the value achieved when the network sees the complete image to a value close to 0.5 AUC ROC when the lungs are completely removed. This drop is not necessarily



(a) Background expansion

(b) Lung exclusion

FIGURE 7. Background expansion and lung exclusion. (a) The original contour area is shown in blue and the expanded area contour in green (b) The contour of the removed area is shown in yellow.

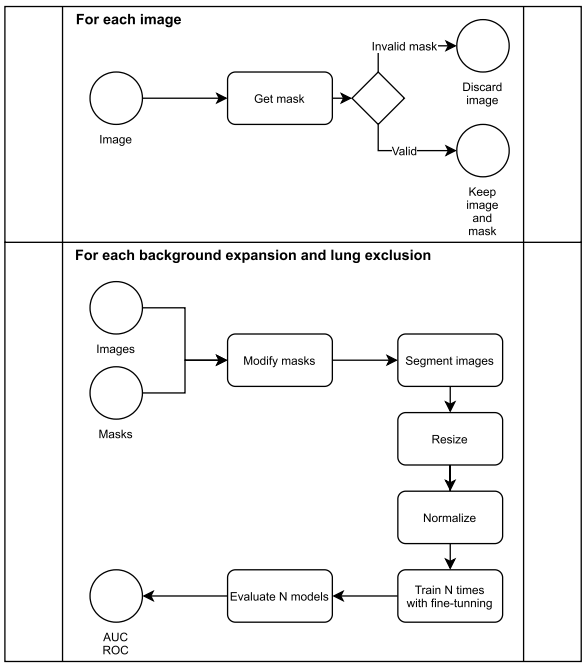


FIGURE 8. Bias analysis's workflow.

linear, but will be shown in the graphs as a straight red line, as can be seen in the right part of Figure 9, to offer a simplified graphical representation of the expected behavior. In the left part of the figure, the green line represents the classification rate obtained using only the lung area.

This analysis has been performed in the three datasets:

- The first one (see Figure 9a), BIMCV, clearly shows a significant bias within the data, as the classification rate steadily increases with the background expansion. The second graph shows that removing the lung area is not associated with a significant decrease in accuracy, as it should, and even with the complete exclusion of the lungs the classifier achieves almost 0.88 AUC ROC.
- The second one (see Figure 9b), RSNA, displays a slightly lower but still consistent bias within the data in both graphs. However, the RSNA dataset was harder to segment than the other ones and, thus, part of the variability shown could arise from poorly segmented images. Nonetheless, a 0.79 AUC ROC is achieved with the lungs completely occluded, which is far from the expected 0.5 AUC ROC.
- The third one (see Figure 9c), CheXpert, conveys interesting results. The left graph's trend is the one expected for an unbiased dataset, as it doesn't vary along with the background expansion. Nevertheless, the precision achieved when the lung is completely occluded is around 0.74 AUC ROC. This implies that the bias is not located specifically in the background, but it must lie in the whole image.

B. COMBINATION STUDY

As mentioned before, the combination study seeks to evaluate how the combination of datasets might provoke the creation of biased data and how the methodology proposed can detect these weaknesses in the final data collection.

The experiments of Section III-A have been reproduced using the combined dataset. Figure 10(a) shows the effect of varying background expansion and lung exclusion when the combination is designed to balance CheXpert with RSNA cases (4878 control and 992 positive pneumonia images from CheXpert plus 3886 positive images from RSNA, giving a balanced dataset with 50% observations from each class).

The last experiment explored a combination of 4878 images of control patients from CheXpert and the whole set of 424 COVID-19 images from COVIDcxr. This dataset combination is typical of the recent crisis scenario, where few images from the new disease are available, they are obtained from different locations, under uncontrolled conditions, with different equipment and acquisition protocols, etc. This is the worst-case scenario and the results are in accordance with it, as can be seen in Figure 10(b).

The results for these experiments show, in a similar fashion to Chexpert's base case, that the bias is ubiquitous in the

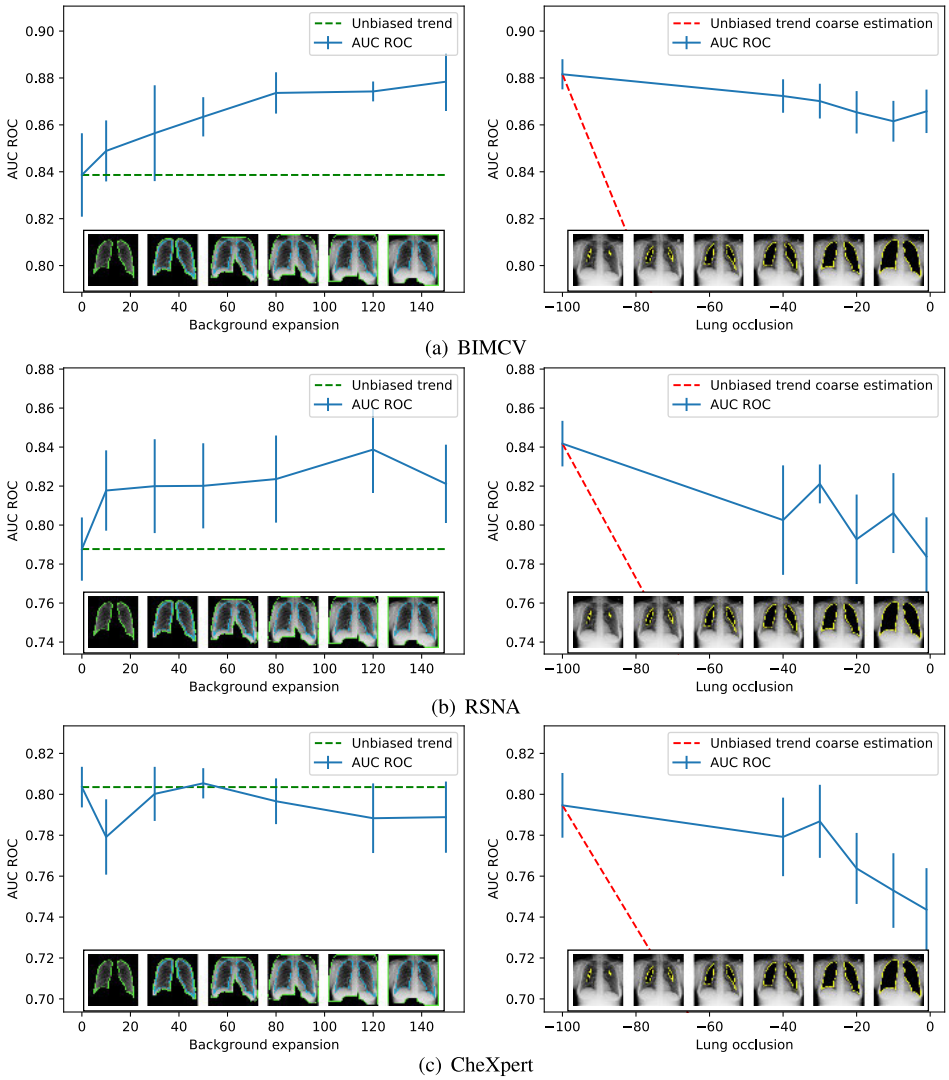


FIGURE 9. Accuracy as a function of the background expansion and lung reduction. The green dotted lines mark the correct behavior of a non-biased dataset when more and more background is included, and the red dotted lines indicate the expected reduction of the classification rate as the lungs are removed from the analysis. Blue lines show the accuracy for a given expansion or reduction with a vertical line indicating the standard deviation.

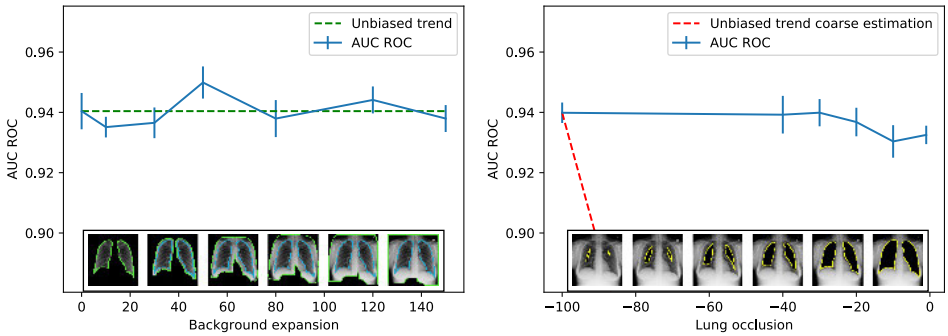
image. Despite increasing the amount of background inside the images doesn't affect the accuracy, the effect of the lung occlusion is not remarkable within the results.

IV. DISCUSSION

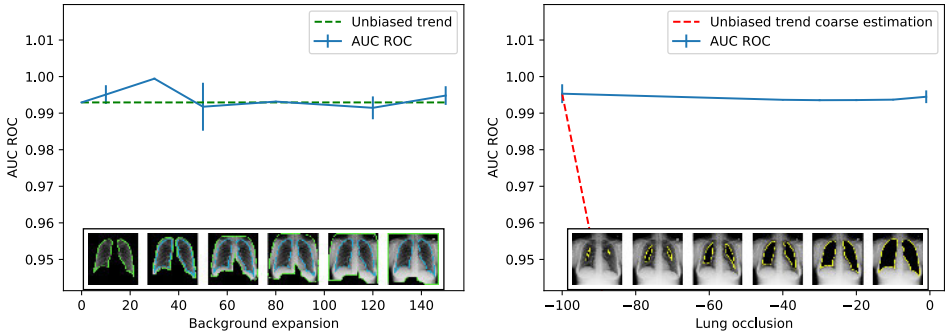
Deep learning has been receiving a lot of attention as a very powerful methodology for analyzing medical images [30]. The ability of Convolutional Neural Networks (CNN) to obtain excellent results even when it is used as a blackbox, as opposed to the classical design of ad-hoc algorithms, has attracted many researchers.

Some works using CNNs for COVID-19 detection on cxr images report high accuracies for a variety of network architectures. In particular, studies using VGG16 report [9] 89.8% accuracy for a dataset built of 180 COVID-19 and 200 control samples, 90% accuracy is obtained [27] for a dataset composed of 202 COVID-19 images, 300 of pneumonia and 300 negative and 93.48% accuracy [31] is achieved using a dataset that contains 224 COVID-19 images, 700 of pneumonia and 504 negative. The fact that VGG16 achieves good results for detecting pulmonary diseases strengthens the hypothesis that the features extracted by the network are relevant to the task and therefore, as detected from our experiments, related to some sort of bias within the images.

One of the drawbacks of CNNs is that they often need large amounts of data to learn and, while generic CXR databases are available, public existing COVID-19 datasets are composed of a few images that were collected by volunteers [16]. As a consequence, these datasets show unbalanced labels and a mix of different data sources that



(a) RSNA positive samples added to CheXpert subset.



(b) COVIDcxr positive samples added to CheXpert subset.

FIGURE 10. Addition of positive samples from RSNA and COVIDcxr to CheXpert's dataset. The green dotted lines mark the correct behavior of a non-biased dataset when more and more background is included, and the red dotted lines indicate the expected reduction of the classification rate as the lungs are removed from the analysis. Blue lines show the accuracy for a given expansion or reduction with a vertical line indicating the standard deviation.

makes getting a robust model and reliable performance measures difficult. In this regard, some articles report the problem of small and unbalanced datasets for COVID-19 detection [4], [32], and propose solutions to mitigate the problem.

Bias analysis has been tackled by other authors. For instance, in [33] the authors proposed that train and test partitions should come from different datasets (related to the same task), as the classifier is trying to achieve maximum performance over a certain task and not over a dataset. This may also assert the true generalization capacity of the classifier. On the other hand, [34] sought to minimize the effects of different biased datasets by way of converting different dataset observations to prototypes, greatly reducing possible intra-dataset specific features.

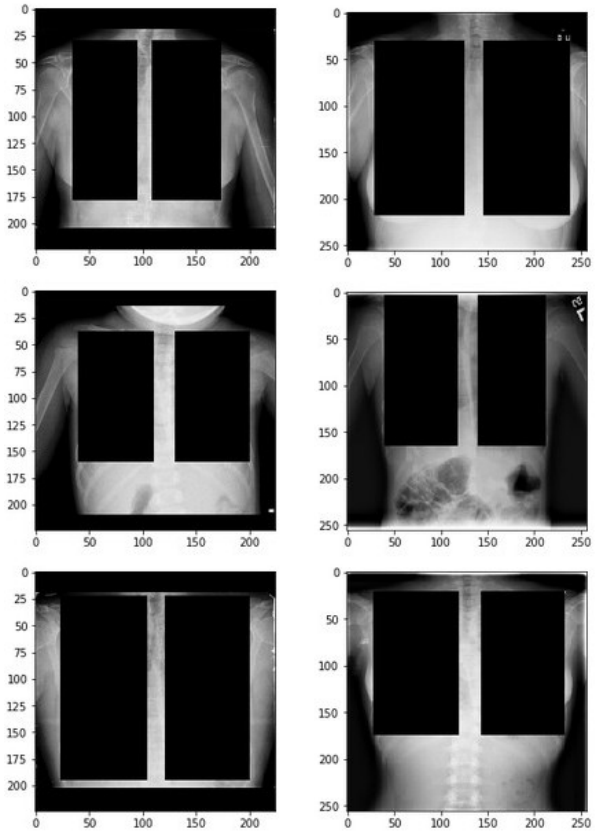
Recently, [35] addresses this issue for COVID-19 detection and reports that the problem of mixing different datasets may lead the network to learn background information. Our study performs a similar approach to the one presented in thisarticle, i.e. both study possible biases within the lungs. [35] occludes the lungs with rectangular fixed-size black boxes and measures the accuracy achieved. However, the proposed methodology extends the concept proposed to more precise masks and progressive inclusion and exclusion of information to the learning process. This allows the ability to detect where the bias approximately is and enables more precise bias estimation.

Furthermore, [36] studies bias within the nCov2019 dataset using information about patients (symptoms, comorbidities, age, and sex). This dataset collects clinical data from different sources rather than images. They found significant bias related to the origin of the data and exposed several issues related to multisource variability.

This article is focused on detecting some biases within widely used CXR datasets to glimpse the degree to which these biases affect the results and proposes a bias detection methodology to assert the validity of results. This methodology makes use of techniques such as heatmap visualization, histogram analysis and selective image occlusion which are combined to evaluate which parts of the images are being used as discriminative features for a classification task. In this work, this methodology has been applied in two case scenarios, one for the existence of bias on individual pneumonia datasets and another to detect the existence of bias in a mix of datasets.

V. LIMITATIONS OF THE STUDY

Regarding possible limitations, there could be a problem with the methodology proposed, since the segmentation masks used for expansion and reduction may be biased themselves. The segmentation process might be more prone to fail in images with pneumonia since the borders of the lungs are more diffuse, whereas this could not happen in images of control patients. This could pose a significant difference



(a) BIMCV (b) RSNA

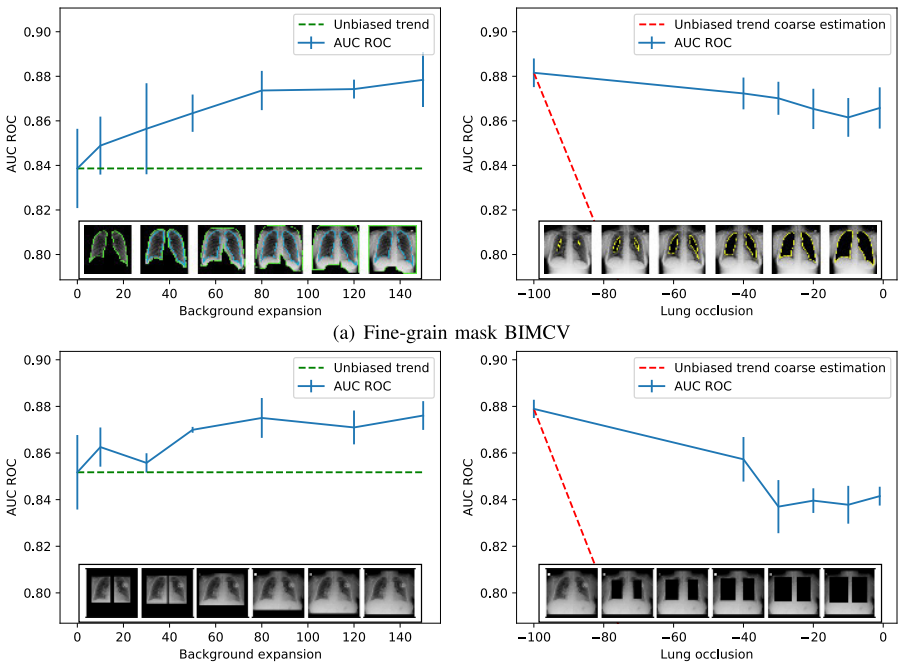
FIGURE 11. Lung occlusion with fixed-size rectangular boxes.

between cases and controls masks, and therefore, we might be introducing a new bias that would imply a problem with the proposed methodology.

However, to rule this out, we designed an experiment where the occlusion masks were substituted by rectangles the size of the lungs. This experiment is similar to the one presented in [35], but here we ensure that the lungs are completely removed using the segmentation mask shape whereas in the aforementioned work they just place a fixed size black rectangle in the central area leaving some lung area uncovered. Some examples from our method can be seen in Figure 11. The results achieved for BIMCV's dataset can be seen in Figure 12, where the differences found are not significant, suggesting that the shape of the lung masks is not influencing the bias detection algorithm proposed.

Furthermore, to increase the confidence in our conclusions, we pre-processed all the images by means of CLAHE histogram normalization to assert how this pre-process affected the results. As can be seen in Figure 13, there is no difference in the results achieved between the normalized and plain images.

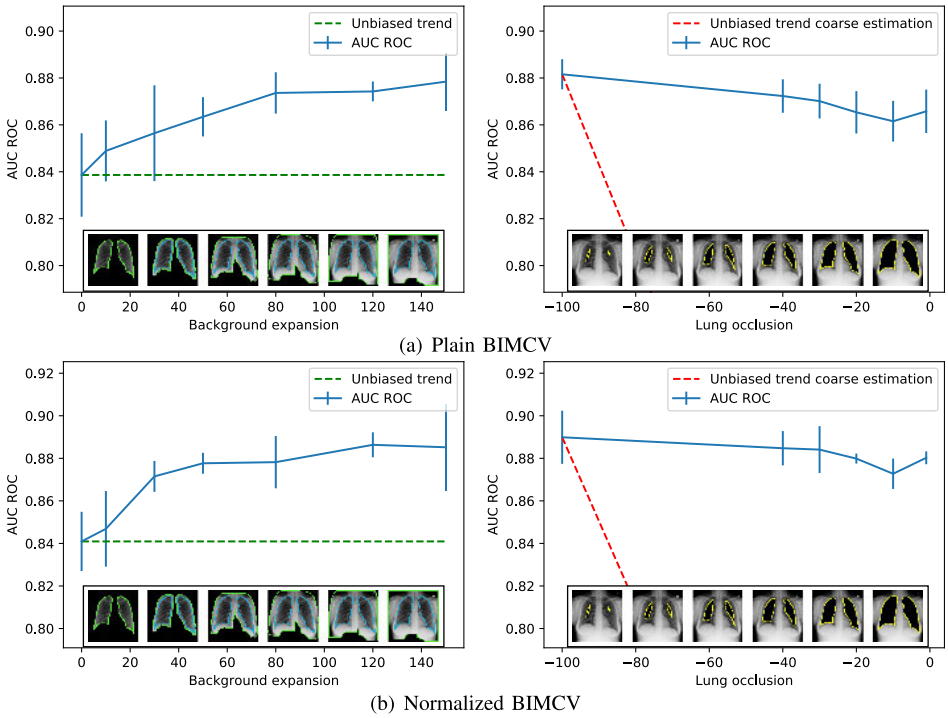
Talking about strengths, the results of the experiments described in Section III-B demonstrated that the classification rate does not improve when the background area is included in the images, which means that either there is no bias specifically on the background or the most significant bias is already within the lungs. However, when the lung area is progressively removed from the image we find in both experiments that the accuracy does not decrease, suggesting



(a) Fine-grain mask BIMCV

(b) Square mask BIMCV

FIGURE 12. Comparison between fine-grain and squared masks for BIMCV's dataset.



(a) Plain BIMCV

(b) Normalized BIMCV

FIGURE 13. **Comparison between normalized and plain BIMCV's dataset.**

that the system is classifying the images according to some elements present in the whole image, not only inside the lungs. That result confirms the hypothesis that powerful systems like Convolutional Networks can find subtle features in the images and give optimistic classification results if no measures are taken to avoid biases in the data.

To summarize, further research should be conducted to reduce the impact of the intrinsic bias for the datasets whose images are collected from several sources. Recent literature has demonstrated the emergence of methodologies useful to reduce the impact of such a bias. Image preprocessing methods [22] or deep learning architectures designed to deal with biased datasets [37] may be a good starting point.

VI. CONCLUSION

In this work, a novel methodology to assess the existence of bias in CXR image datasets is presented. Techniques such as activation heatmap visualization, histogram analysis and selective image occlusion are combined to evaluate which part of the images are being used as discriminative features for a classification task. In this case, the regions of interest were the lungs. The datasets used show different levels of bias, these comprising datasets that try to make information quickly available in an urgent scenario like the current COVID-19 crisis. Some examples are BIMCV's collection or the combination of datasets created for this purpose, which are the ones with more problems. The results are confirmed with the other methodologies used, such as the FDR of the activation map or the histogram analysis.

The study of the effects of combining datasets from different sources is especially interesting because it shows that, if it is not strictly controlled, important biases can be induced in the final dataset. A typical solution for the lack of samples of a given class is to compile different datasets into one that collects all the categories to study, as the recent COVID-19 datasets. In particular, the widely used COVIDcxr dataset, built from different sources, might in fact have included significant biases that inadvertently affected the results published. This kind of heterogeneous dataset often mix observations coming from very diverse equipment, acquisition protocols and processing software. In that context, features found by Deep Convolutional Networks in the images, including the background areas, are enough to get a good classification rate, whilst the actual performance of the classifier for the clinical task attempted can be much lower.

ACKNOWLEDGMENTS

The authors would like to thanks of gratitude to BIMCV and the other teams that compiled and made available the datasets used in this work. The experiments were conducted employing *Instituto Tecnológico de Informática* (ITI) High-Performance Computing platform, which is funded by IVACE and AVI, and implemented within ITI Data Space, being these experiments a TECH4CV's project use case.

REFERENCES

[1] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, J. P. Kanne, S. Raoof, N. W. Schluger, A. Volpi, J. J. Yim, I. B. Martin, and D. J. Anderson, "The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society," *Chest*, vol. 158, no. 1, pp. 106–116, 2020.

[2] M. Rezaul Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, and O. Beyan, "DeepCOVIDExplainer: Explainable COVID-19 diagnosis based on chest X-ray images," 2020, *arXiv:2004.04582*. [Online]. Available: http://arxiv.org/abs/2004.04582

[3] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101794.

[4] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[5] M. Rahimzadeh and A. Attar, "A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of xception and ResNet50 V2," *Informat. Med. Unlocked*, vol. 19, 2020, Art. no. 100360.

[6] M. Polsinelli, L. Cinque, and G. Placidi, "A light CNN for detecting COVID-19 from CT scans of the chest," *Pattern Recognit. Lett.*, vol. 140, pp. 95–100, Dec. 2020.

[7] S. Hosseinzadeh Kassani, P. Hosseinzadeh Kassasni, M. J. Wesolowski, K. A. Schneider, and R. Deters, "Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: A machine learning-based approach," 2020, *arXiv:2004.10641*. [Online]. Available: http://arxiv.org/abs/2004.10641

[8] D. Dansana, R. Kumar, A. Bhattacharjee, D. J. Hemanth, D. Gupta, A. Khanna, and O. Castillo, "Early diagnosis of COVID-19-affected patients based on X-ray and computed tomography images using deep learning algorithm," *Soft Comput.*, pp. 1–9, Aug. 2020, doi: 10.1007/s00500-020-05275-y.

[9] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Syst. Appl.*, vol. 164, Feb. 2021, Art. no. 114054. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417420308198

[10] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: Data augmentation using auxiliary classifier GAN for improved COVID-19 detection," *IEEE Access*, vol. 8, pp. 91916–91923, 2020.

[11] M. Shams, O. Elzeki, M. Abd Elfattah, T. Medhat, and A. E. Hassanien, "Why are generative adversarial networks vital for deep neural networks? A case study on COVID-19 chest X-ray images," in *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*. Cham, Switzerland: Springer, 2020, pp. 147–162.

[12] A. M. Ismael and A. Şengür, "The investigation of multiresolution approaches for chest X-ray image based COVID-19 detection," *Health Inf. Sci. Syst.*, vol. 8, no. 1, pp. 1–11, Dec. 2020.

[13] A. Sarhan. (2020). *Detection of COVID-19 Cases in Chest X-ray Images Using Wavelets and Support Vector Machines*. [Online]. Available: https://europepmc.org/article/PPR/PPR180736

[14] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest X-ray image dataset with multi-label annotated reports," *Med. Image Anal.*, vol. 66, Dec. 2020, Art. no. 101797.

[15] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 590–597.

[16] J. Paul Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," 2020, *arXiv:2003.11597*. [Online]. Available: http://arxiv.org/abs/2003.11597

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[18] C. Sáez, M. Robles, and J. M. García-Gómez, "Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances," *Stat. Methods Med. Res.*, vol. 26, no. 1, pp. 312–336, Feb. 2017.

[19] C. Sáez, O. Zurriaga, J. Pérez-Panadés, I. Melchor, M. Robles, and J. M. García-Gómez, "Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 6, pp. 1085–1095, Nov. 2016.

[20] F. J. Pérez-Benito, C. Sáez, J. A. Conejero, S. Tortajada, B. Valdivieso, and J. M. García-Gómez, "Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0220369.

[21] F. J. Pérez-Benito, J. A. Conejero, C. Sáez, J. M. García-Gómez, E. Navarro-Pardo, L. L. Florencio, and C. Fernández-de-las-Peñas, "Subgrouping factors influencing migraine intensity in women: A semi-automatic methodology based on machine learning and information geometry," *Pain Pract.*, vol. 20, no. 3, pp. 297–309, Mar. 2020.

[22] F. J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, A. Fuster-Baggetto, M. Pollan, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, and R. Llobet, "A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation," *Comput. Methods Programs Biomed.*, vol. 195, Oct. 2020, Art. no. 105668.

[23] M. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*. Cham, Switzerland: Springer, 2008, pp. 315–347.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[25] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[26] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla. (2020). *X-Ray Image Based COVID-19 Detection Using Pre-Trained Deep Learning Models*. [Online]. Available: https://engrxiv.org/wx89s/

[27] T. Zebin and S. Rezvy, "COVID-19 detection and disease progression visualization: Deep learning on chest X-rays for classification and coarse localization," *Appl. Intell.*, vol. 51, no. 2, pp. 1010–1021, 2020, doi: 10.1007/s10489-020-01867-1.

[28] S. Candemir and S. Antani, "A review on lung boundary detection in chest X-rays," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 4, pp. 563–576, Apr. 2019.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[30] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[31] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[32] E. Tartaglione, C. A. Barbano, C. Berzovini, M. Calandri, and M. Grangetto, "Unveiling COVID-19 from CHEST X-ray with deep learning: A hurdles race with small data," *Int. J. Environ. Res. Public Health*, vol. 17, no. 18, p. 6933, Sep. 2020, doi: 10.3390/ijerph17186933.

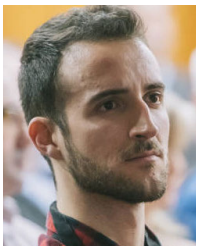
[33] A. Ashraf, S. Khan, N. Bhagwat, M. Chakravarty, and B. Taati, "Learning to unlearn: Building immunity to dataset bias in medical imaging studies," 2018, *arXiv:1812.01716*. [Online]. Available: http://arxiv.org/abs/1812.01716

[34] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.

[35] G. Maguolo and L. Nanni, "A critic evaluation of methods for COVID-19 automatic detection from X-ray images," 2020, *arXiv:2004.12823*. [Online]. Available: http://arxiv.org/abs/2004.12823

[36] C. Sáez, N. Romero, J. A. Conejero, and J. M. García-Gómez, "Potential limitations in COVID-19 machine learning due to data source variability: A case study in the nCov2019 dataset," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 2, pp. 360–364, Feb. 2021, doi: 10.1093/jamia/ocaa258.

[37] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9012–9020.



OMAR DEL TEJO CATALÁ was born in Valencia, Spain, in 1995. He received the master's degree in computer science engineering. He is currently pursuing the Ph.D. degree with the Polytechnic University of Valencia (UPV). He is currently a Computer Science Engineer with UPV. Since 2018, he has been researching with the Instituto Tecnologico de Informatica (ITI), within the Pattern Recognition and Artificial Intelligence Group. He is deeply keen on deep learning techniques applied to several fields such as object detection, medical applications, reinforcement learning, and image classification.

ISMAEL SALVADOR IGUAL received the Advanced Studies Diploma degree in pattern matching, in 2002. Since 2003, he has been working with the Instituto Tecnológico de Informática (ITI) in the Pattern Recognition and Artificial Intelligence Group (PRAIA), where he became a specialist in artificial vision systems for biometrics, medical imaging, and 3-D inspection. He is currently a Computer Science Engineer with the Polytechnic University of Valencia (UPV). He has also led and participated in Research and Development Projects for public institutions as well as for private companies and has published more than 15 articles in the field of machine learning.

FRANCISCO JAVIER PÉREZ-BENITO was born in Salamanca, Spain, in 1988. He received the degree in mathematics from 2006 to 2011 and Technical Engineering in computer systems from 2006 to 2015 from the Universidad de Salamanca, Salamanca, and the Ph.D. degree in mathematics from the Universitat Politècnica de València, Valencia, Spain, in 2020.

He served as a Project Manager for a biotechnology enterprise, Immunostep S.L., from 2012 to 2016, where his interest in research in the biomedical domain arose. From that moment, he collaborated with the Universitat Politècnica de València and, finally, joined the Instituto Tecnológico de la Informática, in 2018. He focused on the characterization of data variability in a clinical environment and how this variability may influence the machine learning models' performance. These interests drove the publication of several scientific articles in highly cited journals. The topics of his scientific contributions mainly cover applied mathematics, computer science, and artificial intelligence.

DAVID MILLÁN ESCRIVÁ studied computer engineering and received the master's degree in computer vision, artificial intelligence, and computer graphics from the Polytechnic University of Valencia (UPV). He has worked in some startups and companies like Skin Analytics Ltd., as a Computer Vision and Machine Learning Team Leader, and as a Web Developer with Artres Comunicación. He has also worked with the Emotion Research Laboratory in Machine Learning and Computer Vision fields for three years. He has published several books about OpenCV development. Since November 2018, he has been with Instituto Tecnológico de Informática (ITI) in the fields of computer vision, machine learning, and software engineering.

VICENT ORTIZ CASTELLÓ was born in Oliva, València, Spain, in 1991. He received the B.S. degree and the M.S. degree in industrial engineering from the Universitat Politècnica de València, València, in 2015. He is currently pursuing the B.S. degree in telecommunications engineering with the Universitat Oberta de Catalunya, Barcelona, Spain, and the M.S. degree in artificial intelligence with the Universidad Internacional Menéndez Pelayo, Santander. In 2015, he was a Research Assistant with the Instituto de Automática e Informática Industrial, Universitat Politècnica de València, València. From 2016 to 2018, he was a Researcher in biomechanics with the Instituto de Biomecánica de València, Universitat Politècnica de València. Since 2018, he has been a Researcher in artificial intelligence and computer vision with the Instituto Tecnológico de Informática at Universitat Politècnica de València.

RAFAEL LLOBET received the Ph.D. degree in computer science from the Universitat Politècnica de València (UPV), Spain, in 2006.

He has worked with the Instituto de Biomecánica de Valencia (IBV) and with the Instituto Tecnológico de Informática (ITI). Since 2000, he has been with UPV, where he serves as an Assistant Lecturer with the Department of Information Systems and Computation. He also collaborates with ITI where he develops his research. He has published works in 14 international journals and 13 international conferences. His current research interests include machine learning and its application to healthcare area. His research is mainly focused on medical image processing, genomics data analysis, and computer-aided diagnosis.

JUAN-CARLOS PERÉZ-CORTÉS received the Ph.D. degree in computer science from the Polytechnic University of Valencia. He is currently a Full Professor. He is also the Director of the Pattern Recognition and Image Analysis (PRAIA) Research Group with the Instituto Tecnológico de Informática (ITI). He has led and coordinated research projects funded by public national and international entities in the field of medical imaging, industrial software, computer vision, pattern recognition, free software, and so on. He teaches master's degree and Ph.D. courses in computer system's artificial vision and pattern recognition. He has published works in 15 journals, three books, and 17 books and conferences, and has been awarded by public and private entities.