# IMPACT OF LUNG SEGMENTATION ON THE DIAGNOSIS AND EXPLANATION OF COVID-19 IN CHEST X-RAY IMAGES

A PREPRINT

**Lucas O. Teixeira**
Universidade Estadual de Maringá
Maringá, PR, Brazil
lucasxteixeira@gmail.com

**Rodolfo M. Pereira**
Instituto Federal do Paraná
Pinhais, PR, Brazil
rodolfomp123@gmail.com

**Diego Bertolini**
Universidade Tecnológica
Federal do Paraná
Campo Mourão, PR, Brazil
diegobertolini@gmail.com

**Luiz S. Oliveira**
Universidade Federal do Paraná
Curitiba, PR, Brazil

**Loris Nanni**
Università Degli Studi di Padova
Padova, Italy

**Yandre M. G. Costa**
Universidade Estadual de Maringá
Maringá, PR, Brazil
yandre@din.uem.br

September 22, 2020

## ABSTRACT

The COVID-19 pandemic is undoubtedly one of the biggest public health crises our society has ever faced. The impact of COVID-19 on our society, economy, and healthcare system is estimated to be high, given its clinical severity. Timely and accurate diagnosis is crucial to allow the healthcare personnel to take the appropriate actions. One of the main complications caused by COVID-19 is pneumonia. The standard exams for pneumonia diagnosis are chest X-ray (CXR) and computed tomography (CT) scan. The CT scan is more precise than the CXR. However, CXR is quite suitable in some particular situations because it is cheaper, faster, more widespread in less economically developed regions, and exposes the patient to much less radiation. This paper's main objectives are to demonstrate the impact of lung segmentation in COVID-19 automatic identification using CXR images and evaluate which contents of the image decisively contribute to the identification. We have performed lung segmentation using a U-Net CNN architecture, and the classification using three well-known CNN architectures: VGG, ResNet, and Inception. To estimate the impact of lung segmentation, we applied some Explainable Artificial Intelligence (XAI), such as LIME and Grad-CAM. To evaluate our approach, we built a database named RYDLS-20-v2, following our previous publication and the COVIDx database guidelines. We evaluated the impact of creating a COVID-19 CXR image database from different sources, called database bias, and the COVID-19 generalization from one database to another, representing our less biased scenario. The experimental results of the segmentation achieved a Jaccard distance of 0.034 and a Dice coefficient of 0.982. In the best and more realistic scenario, we achieved an F1-Score of 0.74 and an area under the ROC curve of 0.9 for COVID-19 identification using segmented CXR images. Further testing and XAI techniques suggest that segmented CXR images represent a much more realistic and less biased performance. More importantly, the experiments conducted show that even after segmentation, there is a strong bias introduced by underlying factors from the data sources, and more efforts regarding the creation of a more significant and comprehensive database still need to be done.

***Keywords*** COVID-19 · chest X-ray · semantic segmentation · explainable artificial intelligence.

# 1 Introduction

The Coronavirus disease 2019 (COVID-19) pandemic, caused by the virus named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has become the most significant public health crisis our society has faced recently [1]. COVID-19 affects mainly the respiratory system and, in extreme cases, causes a massive inflammatory response that reduces the total lung capacity [2].

In such extreme cases, the patient needs constant hospital care, possibly in an Intensive Care Unit (ICU), to use a mechanical ventilator to help to breathe and avoid hypoxia (low blood-oxygen levels) [3]. That characteristic associated with high transmissibility, lack of general population immunization, and high incubation period [4] makes COVID-19 a dangerous and lethal disease. In these circumstances, artificial intelligence (AI) based solutions are being used in various contexts, from diagnostic support to vaccine development [5].

The standard imaging tests for pneumonia, and consequently COVID-19, are chest X-ray (CXR) and computed tomography or computerized X-ray imaging (CT) scan. The CT scan is the gold standard for lung disease diagnosis since it generates images with a large variety of details. However, CXR is still very useful in these scenarios, since they are cheaper, generate the resulting images faster, expose the patient to much less radiation, and it is more widespread in the emergency care units [6].

After the COVID-19 outbreak, several studies were proposed to investigate its diagnostic based on the use of images taken from the lungs [7, 8, 9, 10, 11]. Despite the impressive advances, there is a lack of more critical analysis regarding the content captured in those images that contribute to the achievement of consistent results [12, 13, 14]. As a brief demonstration of this lack of critical analysis, a quick search on Google Scholar[1], looking for works on the subject of COVID-19 identification in CXR that do not mention the segmentation technique, returns a total of 2.760 results.

Our main objective is to evaluate the impact of lung segmentation in identifying pneumonia caused by different microorganisms using CXR images obtained from various sources (i.e., Cohen, RSNA pneumonia detection challenge, among others). We have primarily focused on CXR images due to its smaller cost and high availability in the emergency care units, especially those located in less economically developed regions. Moreover, we put more emphasis on COVID-19, aiming to provide solutions that can be useful in the current pandemic context. To support that objective, we used an U-Net Convolutional Neural Network (CNN) for lung segmentation, and three popular CNN models for COVID-19 identification: VGG16 [15], ResNet50V2 [16] and InceptionV3 [17].

The first step towards that objective was to improve our previously created COVID-19 database (i.e., RYDLS-20 [9]), renamed as RYDLS-20-v2. We have added more data sources for this expansion and have gathered more images from the previously used sources. Furthermore, we adopted a similar process employed by the COVIDx dataset [11], which is probably the most used COVID-19 CXR image dataset.

Firstly, we have designed the problem as a multi-class classification problem with three classes: lung opacity, COVID-19, and normal lungs (i.e., no-pneumonia), in which lung opacity means pneumonia caused by any previously known pathogen. Secondly, we have experimentally tested a classification scenario with more granular labels, such as pneumonia caused by different viruses, bacterias, fungi, among others. However, in nearly all scenarios, the models did not correctly differentiate lung opacity types well enough. This is also consistent with clinical practice in which it is difficult to differentiate between viral and bacterial pneumonia [18, 19]. Furthermore, in order to provide a more complete and realistic overview, we also evaluated specific scenarios to assess the database bias, i.e., the importance of the image source for the classification model and COVID-19 generalization, i.e., the usage of COVID-19 images from one database to train a classification model to identify COVID-19 cases in a different database, which represents the less biased scenario evaluated in this paper.

To achieve a satisfactory lung segmentation, we experimented with some well-known techniques such as thresholding, edge detection, watershed, and deep learning based strategies [20]. The results obtained using classical approaches (thresholding, edge detection, and watershed) were not even near satisfactory, probably because a CXR image is very homogeneous, and such algorithms depend on the contrast occurrence to properly perform the segmentation. Additionally, since we have included images from different data sources with probably different imaging protocols, X-ray machines, etc., the CXR images itself may differ from each other, which makes static segmentation approaches ineffective. Finally, for the deep learning approach, we have used a U-Net CNN architecture [21]. The U-Net has two main components: a contraction path that reduces the input dimension to an encoding and an expansion path that increases the encoding to match the original input dimension. The activation maps in the contraction path are also copied to the expansion path to retain as much information as possible.

[1] We have performed this search on September $1_{st}$, 2020 and have used the following search key: “COVID-19” “X-ray” OR “CXR” “machine learning” OR “artificial intelligence” -segmentation

As known, a deep learning approach is "data-hungry"; i.e., it performs better as the training data increases [22]. Therefore, specifically for the segmentation task, we also used three additional datasets containing binary lung masks to increase the training data: Montgomery County X-ray Set [23], Shenzhen Hospital X-ray Set [23, 24] and Japanese Society of Radiological Technology (JSRT) [25].

Moreover, we adopted an iterative workflow for lung segmentation with the following steps: i) train the U-Net CNN model from scratch; ii) predict binary masks for all CXR images in our COVID-19 database; iii) manually review all predicted masks; and iv) manually create binary masks for wrong predictions. We repeated this process until we arrived at an acceptable model. We included every visible region of both lungs for the manually created masks, usually until the diaphragm extent and most of the heart region. The heart's inclusion is beneficial for two reasons: opacities behind it might be relevant for viral pneumonia, and uniform shaped lungs reduce the impact of the heart size from the left lung [26].

Over the last few years, the area known as Explainable Artificial Intelligence (XAI) has attracted many researchers in the artificial intelligence (AI) field. The main interest of XAI is to research and develop approaches to explain the individual predictions of modern machine learning (ML) based solutions. The rationale behind these methods is to look for a reasonable explanation about why an AI technique achieves a specific performance on a given task, and not just what is the performance. In medical applications based on images, in particular, we understand that a proper explanation regarding the obtained decision is fundamental. In an ideal scenario, the decision support system should be able to suggest the diagnosis and show, as better as possible, which contents of the image, and from which parts, have decisively contributed to achieving a particular decision.

XAI methods can be useful in many different perspectives, considering the general framework widely used to develop pattern classifier systems. Regarding the approach used to obtain the representations (i.e., handcrafted vs. non-handcrafted features), XAI could help to open the "black box", allowing to understand better which specific contents of the image captured are contributing in a definitive way to the decision taken, and how they have been captured. In another vein, as is known, depending on the algorithm used to create the classifier model, how the decision was achieved can be easier or harder to understand. Again, XAI methods can also be helpful in this situation.

To assess the impact of lung segmentation on the identification of COVID-19, we used two XAI approaches: Local Interpretable Model-agnostic Explanations (LIME) [27] and Gradient-weighted Class Activation Mapping (Grad-CAM) [28]. LIME works by finding features, superpixels (i.e., particular zones of the image), that increases the probability of the predicted class, i.e., regions that support the current model prediction. Such regions can be seen as important regions because the model actively uses them to make predictions. LIME is model-agnostic, hence it does not need any details about the model structure or algorithm; i.e., it does not rely on the model to produce the explanations. Thus, the prediction explanation is more unbiased towards the model. On the other hand, Grad-CAM focuses on the gradients flowing into the last convolutional layer of a given CNN for a specific input image and label. We can then visually inspect the activation mapping (AM) to verify if the activations are focusing on the appropriate portion of the input image. Grad-CAM can be applied to various CNN models, however as it depends on the activations of the last convolutional layer to produce the output, it is model-specific. Thus, in a way, both techniques are complementary, and by exploring them, we can provide a complete report of the lung segmentation impact on COVID-19 identification.

The lung segmentation might not improve the classification performance metric itself, because when the whole image is considered, the model may learn to use other features besides lung opacities, or even from outside the lungs region. In such cases, the model is not learning to identify pneumonia or COVID-19, but something else. Thus, we can infer that the model is not reliable even though it achieves a good classification performance. Using lung segmentation, we would supposedly remove a meaningful part of noise and background information, forcing the model to take into account only information from the lung area, i.e., desired information in this specific context. Thus, the classification performance in models using segmented CXR images tends to be more realistic, closer to human performance and better reasoned.

The remaining of this paper is organized as follows: Section 2 introduces some basic concepts used throughout this study. Then, Section 3 presents some related works and how they are related to our paper. After that, Section 4 introduces our proposed methodology. Section 5 shows details about our experimental setup, including database, algorithms, and parameters. Section 6 presents the obtained results. Later, Section 7 discusses the obtained results. Finally, Section 8 presents our conclusions and possibilities for future works.

## 2 Theoretical Background

This section briefly discusses concepts regarding pneumonia and COVID-19, deep learning, and explainable artificial intelligence (XAI) that are important to understand the remainder of this paper.

### 2.1 COVID-19

The first COVID-19 case was reported in Wuhan, China, at the end of 2019. It rapidly evolved from a local epidemic to a global pandemic in a matter of three months. In March 2020, there were cases reported in almost every country in the World, and most of them were applying measures to contain the virus spread, such as social distancing, use of face masks, and constant decontamination [29]. As of right now, there are approximately 21 million confirmed cases worldwide by the World Health Organization (WHO) [1].

The impact of COVID-19 depends upon three factors: the number of people currently infected, virus transmissibility, and clinical severity [30]. Even today, the actual fatality rate cannot be precisely estimated. Among patients that were medically attended, the fatality rate is estimated at around 2% [31]. Early studies estimated the transmissibility, also known as basic reproduction number or $R_0$, to be between 2.2 and 2.5 [32], which means that each infected person will pass it to two other people; this number has effectively gone down after the social distancing measures were adopted. The number of people currently infected has been stable lately, which is the first sign of a plateau in the outbreak.

COVID-19 affects primarily the respiratory system causing, in some cases, pneumonia. Pneumonia is an inflammation in the lungs affecting the oxygen transfer. It cannot be classified as a unique disease, but as a group of different diseases with different characteristics depending upon the infectious agent that caused the inflammatory response. In the case of COVID-19, image tests revealed multiple peripheral ground-glass opacities in the subpleural region of the lungs [31].

The primary image diagnosis for pneumonia, in general, are chest X-rays (CXR) and computed tomography (CT) scan. CT scan produces better and more precise images than CXR. However, the CXR is still useful in some situations, and recent studies showed that it is possible to screen patients with early symptoms based on CXR images [33]. The CXR is broadly available in the emergency care units, especially in low-income regions, have a speedy turnaround time, requires much less decontamination between patients, and less radiation exposure, which means that the test can be performed multiple times to assess the disease evolution. The pneumonia diagnosis from CXR images is not straightforward, and in some cases, even experienced medical practitioners face difficulties [34].

In the COVID-19 research context, AI and pattern recognition techniques have been actively tested and used with different purposes, from diagnostic support to vaccine development [5].

### 2.2 Deep learning

The application of deep learning approaches has recently gained much traction, mainly because of the significant increase in the computational power available. Despite that, approaches that used the backpropagation algorithm and automatic differentiation had been proposed a long time ago in the literature [35].

Deep learning is part of a broader set of machine learning algorithms based on artificial neural networks (ANN). An ANN is composed of multiple smaller units called neurons. Each neuron is a processing element that receives inputs and produces outputs based on an activation function. The combination of multiple neurons in multiple layers produces a compelling computational model capable of representing very complex problems [36]. Furthermore, the term deep refers to a multi-layer ANN. The minimum number of layers that characterize a deep ANN is not very clear in the literature.

The field has dramatically advanced in recent years, and now we have many specialized operations for specific applications. One of them is the convolutional neural network (CNN) for image analysis, classification, and segmentation. The CNN resembles the organization of the human visual cortex [37, 38].

There have been several CNN architectures proposed over the years for many different problems. This work has used three popular CNN architectures taken from the literature: VGG16, ResNet50V2, and InceptionV3. Furthermore, we adopted the framework Keras[2] on top of TensorFlow[3] for our experiments. Keras is the high-level API of TensorFlow (TF): a very intuitive and productive interface to TF focused on deep learning [39]. Furthermore, we also used transfer learning in all models by loading weights trained on ImageNet[4]. Transfer learning is the usage of weights from a CNN trained on a different and larger dataset. The objective is to leverage the knowledge (features) already obtained from a different dataset [40]. ImageNet is a huge and broad dataset that is commonly used as the standard image dataset benchmark.

[2]https://keras.io/

[3]https://www.tensorflow.org/federated

[4]The ImageNet project is a large visual database designed for use in visual object recognition software research. Available at http://www.image-net.org/

### 2.3 Explainable AI

As the methods and algorithms in ML evolved, many of them became highly complex and obscure, especially deep learning models. Thus, it became tough for a human to comprehend how these models work when making predictions. Such models are frequently said to be black-box classification models.

Explainable AI is an area focused on methods and approaches that can be used to explain the model predictions. The main idea is to identify what features the model is actively using when making predictions. Especially when considering complex models, such as CNNs, there is no guarantee of which feature the model will focus on when learning. It will always try to increase the classification metric provided by all means necessary. For images, for instance, it might be not using the relevant portion of the image containing the appropriate information, but something else, given that it increases the classification performance. Hence, model inspection is vital to guarantee that our model uses the appropriate information for prediction [41].

In this paper, we applied two XAI approaches to evaluate how the segmentation impacts the classification process: Local Interpretable Model-agnostic Explanations (LIME) [27] and Gradient-weighted Class Activation Mapping (Grad-CAM) [28]. They both aim to explain individual predictions by finding important portions of the input image that are actively used by the model. Nonetheless, they differ on a fundamental characteristic: LIME is model-agnostic, it does not rely on any detail of the classification model apart from the prediction, while Grad-CAM is model-specific, it uses the activation mapping from the last convolutional layer of a given CNN. Thus, by leveraging both approaches, we hope to present a complete result.

LIME creates a new dataset containing a sequence of examples based on a specific input image with small changes in its features and the black-box model prediction and creates a sparse linear model around them, and then it estimates what features increase or decreases the predicted probability of a given class [27]. The sparse linear model weights the examples by the degree of change from the original input. Such regions can be understood as supporting and contradicting regions, despite that these regions are important and are being actively used during predictions.

On the other hand, Grad-CAM can only be used on CNN models and leverages the CNN structure to find the important regions. The idea behind Grad-CAM is quite intuitive: it uses the activation mapping, also called feature map or simply activations, of the last convolutional layer and weights them by the predicted probability of a given label. The regions where the activation mapping is large are the regions where the final prediction uses the most.

Finally, we are using LIME and Grad-CAM to assess whether the models using segmented images focus primarily on lung area information, which might not be the case for a model using full CXR images.

## 3 Related Works

This section discusses some noteworthy papers found in the literature related to one of the following topics: lung segmentation in CXR and CT images and COVID-19 model inspection and explainability. We present the main aspects, including experimental database, type of image (CXR or CT), segmentation, and classification model.

[42] proposed a deep learning approach for infection segmentation on a CT scan using a VB-Net architecture. The VB-Net is a combination of the V-Net with the bottle-neck structure. The dataset comprised 249 CT scan images for training and 300 CT images for validation, all of them diagnosed with COVID-19. It adopted a human-in-the-loop strategy for the manual contouring of the infections, i.e., the human reviewed and improved the automated prediction for the training data. Their approach for segmentation of infection regions presented a Dice similarity coefficients of $91.6\% \pm 10.0\%$.

[43] proposed a framework to detect COVID-19 through 3D chest CT images. The dataset was composed of CT images taken from 3322 patients, collected between August 2019 and February 2020. The authors describe rates with 0.96 of AUC for COVID-19, and 0.98 of AUC for Non-Pneumonia. This work used a COVNet framework that employs RESNET-50 for feature extraction and U-net as a segmentation method.

[44] developed residual attention U-Net model. The model uses ResNeXt blocks in the encoder and decoder paths. The encoder path also uses an attention mechanism to retain important information for the correct segmentation. The dataset used was the COVID-19 CT Segmentation dataset, provided by the Italian Society of Medical and Interventional Radiology (SIRM), which contains 110 axial CT images and 100 segmentation masks[5,6]. They set up the segmentation as a multi-class problem with three labels: ground-glass opacity, consolidations, and pleural effusion. The results showed a Dice similarity coefficient of 94%.

[5] https://www.sirm.org/category/senza-categoria/COVID-19/

[6] http://medicalsegmentation.com/covid19/

[45] presented a Joint Classification and Segmentation (JCS) system to perform explainable COVID-19 classification. The classification is performed using a Res2Net CNN architecture, and the segmentation is performed using VGG16 CNN as a backbone together with a Grouped Atrous Module (GAM) and an Attentive Feature Fusion (AFF). The explanation is performed by extracting the convolutional layers activation mappings. The proposal was evaluated on a dataset composed of 144,167 CT scan images from 400 COVID-19 positive examples and 350 negative patients, all confirmed by RT-PCR tests. The results showed an average sensitivity of 95.0% and a specificity of 93.0% on the classification task, and 78.3% Dice score on the segmentation task of infection regions.

[46] proposed a small and efficient segmentation architecture for COVID-19 CT scan images, named MiniSeg. They proposed a new CNN block called Attentive Hierarchical Spatial Pyramid (AHSP) to perform the segmentation. The dataset used was the COVID-19 CT segmentation dataset, which contains 100 axial CT images with three segmentation classes: ground-glass opacity and pleural effusion. Their results presented an average IoU of 82.12%.

[47] proposed a deep neural network, called COVID-SegNet, aiming at segmenting the infection spots and the entire lung region from CT scan images. They proposed a new block named Feature Variation (FV), specifically designed to identify and adapt to a global context to identify the infection spots correctly. The network also fuses features at different scales by using a method called Progressive Atrous Spatial Pyramid Pooling (PASPP) proposed by the authors, in order to be able to identify different shapes of the infections spots. The database was composed of 21,658 annotated chest CT images from 861 patients with confirmed COVID-19. The results showed a Dice similarity coefficient of 98.7% and 72.6% for lung and COVID-19 infection regions segmentation.

[11] proposed a specific CNN design for COVID-19 identification in CXR images, named COVID-Net. The COVID-Net contains various blocks, such as projection-expansion-projection-extension (PEPX), convolutions, and long-range connections. They built a comprehensive COVID-19 CXR dataset, called COVIDx, containing 16,756 images with three classes: pneumonia, COVID-19, and normal. The results showed a 92.4% of accuracy for the COVIDx test.

[48] proposed a novel COVID-19 lung CT infection segmentation network, named Inf-Net. They also proposed a semi-supervised approach to increase the performance given the low sample size of COVID-19 CT scan images, named Semi-Inf-Net. The Inf-Net adopts a parallel partial decoder (PPD) to aggregate features from high-level layers, recurrent reverse attention (RA) modules, and explicit edge-attention to improve the boundaries. The dataset used was the COVID-19 CT Segmentation dataset, provided by the Italian Society of Medical and Interventional Radiology (SIRM), already mentioned here. They also used unlabelled examples in the COVID-19 CT Collection [49]. They set up the segmentation as a multi-class problem with two labels: ground-glass opacity and consolidation. The results showed a Dice similarity coefficients of 68.2% and 73.9% for Inf-net and Semi-Inf-Net, respectively, for the infection region segmentation, and an average 47.4% and 54.1% for Inf-net and Semi-Inf-Net for the multi-class segmentation, respectively.

[50] presented an approach for COVID-19 CXR classification and explanation using Grad-CAM, Grad-CAM++ and LRP (layer-wise relevance propagation). The dataset was based on COVIDx [11] with some additions resulting in 15,959 CXR images. They explored some popular CNN architectures, such as VGG, ResNet, and DenseNet. Furthermore, they also tried some ensembles. The best result is an F1-Score of 0.945 for the multi-class setup. For some examples, they explored the explainability investigating multiple approaches: Grad-CAM, Grad-CAM++, and LRP.

[51] performed the discrimination between viral pneumonia and non-viral pneumonia using Confidence-Aware Anomaly Detection (CAAD) on Chest X-ray Images. The CAAD approach has three steps: Shared feature extractor, an anomaly detection step, and a confidence prediction module. The experiments were carried out on two databases (X-VIRAL and X-COVID). Experiments using the X-COVID dataset, composed of 106 positive COVID-19 images, and 107 normal samples, describe an AUC rate of 83.61% and a sensitivity of 71.70%. According to the authors, the rates are compared to the performance of radiologists.

[52] applied a series of CNN models for COVID-19 identification considering a binary and a multi-class scenario and using weighting and oversampling to overcome the dataset balance issue. The dataset used contains 1214 CXR images, of which 108 are COVID-19 samples. The best result is an F1-Score of 0.98 and 0.96 for the binary and multi-class scenarios, respectively. For a couple of images, they explored the explainability investigating multiple layers activations, class activation maps (CAM), and LIME.

[53] presented a very detailed paper exploring classification and explanation using CXR and CT scan images. The classification achieved an F1-Score of 0.9 with CT scans and 1 with CXR, both using NasNetMobile CNN architecture. We must highlight that they also tested other CNN architectures, and overall the models using CXR images surpassed the models using CT scan images, which is an exciting result since we expected that CT scan images would be better. The CXR dataset was composed of 400 CXR(200 COVID-19 cases and 200 Non-COVID-19 cases). The CT scan dataset followed precisely the same distribution. Besides, the authors investigated multiple layers activations on an XAI perspective, using LIME.

[54] accomplished the estimation of COVID-19 risk by analyzing CXR images. In the proposal, the authors created a pipeline in which the first step is to perform the lungs' segmentation. The opacity regions, very peculiar in CXR affected by pneumonia, were considered as missing data. A modified CNN for image segmentation based on a deep generative model for data imputation was presented. In conclusion, the authors advocate that the segmentations' quality is sufficient to use it to score COVID-19 risk.

[55] used a dual sampling attention network to diagnose COVID-19 from community-acquired pneumonia using chest CT images. The authors developed a novel schema aiming to perform the diagnose focusing on the infection regions. They also adopted a strategy to deal with imbalance issues, and they claim that the proposal was evaluated on the most considerable multi-center CT data for COVID-19. A total of 2186 CT scans taken from 1588 patients were used with 5-fold cross-validation for training-validation purposes. On the other hand, the test was done on an independent dataset composed of 2796 CT images obtained from 2057 patients. The results obtained for COVID-19 identification achieved 0.944 of AUC.

Despite the novelty of COVID-19 as a research topic, we have witnessed an impressively huge number of works that have been launched every day. In this vein, [56] carried out a "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19". The authors focused both on CT and X-ray images, and the entire pipeline of medical imaging and analysis techniques involved with COVID-19 was covered. In conclusion, the authors highlight that even though imaging techniques empowered by AI have been proven to be successful for COVID-19 detection, we must not forget that these techniques provide only partial information. The clinical manifestations and laboratory examination cannot be ignored.

Due to the novelty of the COVID-19 pandemic, new papers are emerging every day. Thus it is unfeasible to show an accurate state-of-the-art. Moreover, many studies consider different datasets, which makes direct comparisons challenging or not even possible.

Finally, from this brief literature overview, we can figure out that most papers focused primarily on deep learning approaches applied to CT scan images, which is somehow different from our proposal since we are dealing with CXR images. Furthermore, the papers that investigated XAI only showed a few examples of explanations and did not generalized them for all images. The problem is that such examples may have been handpicked to show precisely the expected behavior the authors wanted to. Besides, as stated in the introduction section, there is a massive number of works in the literature that performs COVID-19 identification in CXR images that do not even mention the segmentation technique.

## 4 Proposed Method

As previously mentioned, we focus on exploring data from CXR images for reliable identification of COVID-19 among pneumonia caused by other micro-organisms. Hence, we proposed a specific method that allowed us to assess lung segmentation's impact on COVID-19 identification reliably.

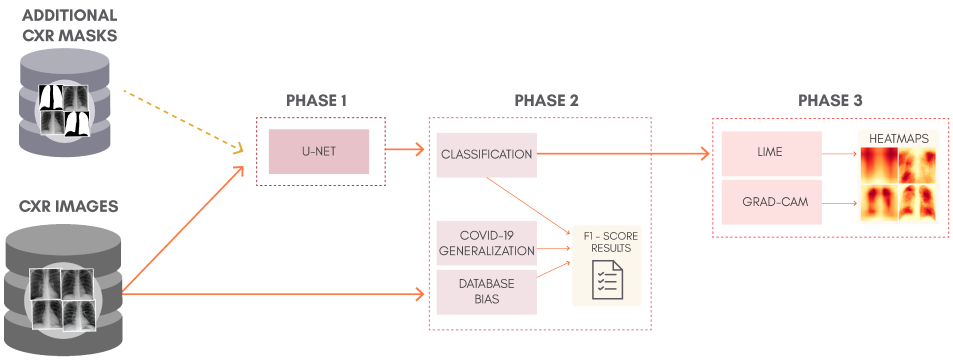



Figure 1: Proposed methodology.

To better understand the proposal of this work, Figure 1 shows a general overview of the classification approach adopted, containing: the lung segmentation (Phase 1), classification (Phase 2), and prediction explanation (Phase 3). In Figure 1, Phase 1 is skipped entirely for the classification of non-segmented CXR images.

It is essential to point out that our dataset contains CXR from different sources, and they are not standardized at all. Thus, we are dealing with images from different X-ray machines and operators, leading to very different protocols affecting the resulting image. The dataset will be described in more detail in section 5.2.

In the following subsections, we describe each one of the Phases described in Figure 1.

### 4.1 Lung Segmentation (Phase 1)

The first phase in our method is the lung segmentation, aiming to remove all background and retain only the lung area. By removing the background, we expect to reduce noise that can interfere in the model prediction. Figure 2 presents an example of lung segmentation.

(a) CXR image. (b) Binary mask. (c) Segmented lungs.

Figure 2: Lungs segmentation on CXR image.

Specifically, in deep models, any extra information can lead to model overfitting. This is especially important in CXR since many images contain burned-in annotations about the machine, operator, hospital, or patient. Figure 3 presents an example of CXR images with burned in information.

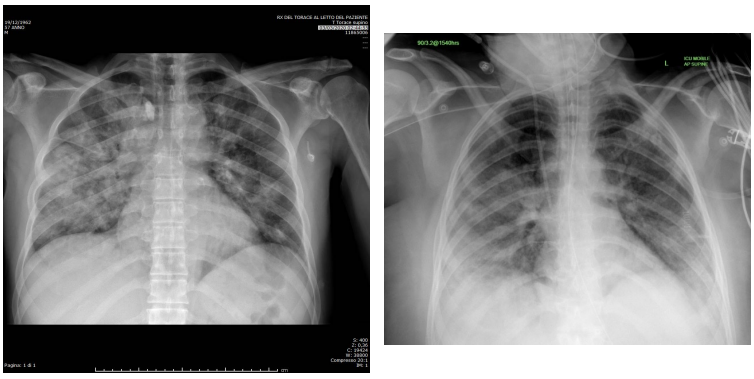


Figure 3: CXR with burned in annotations.

Our primary purpose is not necessary to achieve an improvement in the classification performance metric. However, we expect an increase in the model reliability and prediction quality in a real-world scenario. We expect that the models using segmented images rely on information in the lung area rather than background information. For example, if a model is trained to predict lung opacity, it must use lung area information. Otherwise, it is not predicting opacity but something else.

In order to perform lung segmentation, we applied a CNN approach using the U-Net architecture [21]. The U-Net input is the CXR image, and the output is a binary mask that indicates the region of interest (ROI). Thus, the training requires a previously set of binary masks.

The COVID-19 dataset used does not have manually created binary masks for all images. Thus, we adopted a semi-automated approach to creating binary masks for all CXR images. First, we used three additional CXR datasets with binary masks to increase the training sample size and some binary masks provided by v7labs[7]. We then trained the U-Net model and used it to predict the binary masks for all images in our dataset. After that, we reviewed all predicted binary masks and manually created masks for those CXR images, which the model was unable to generalize well. We repeated this process until we judged the result satisfactory and achieved a good intersection between target and obtained regions. It is important to note that we also applied many data augmentation transformations to extend our training data further.

#### 4.1.1 U-Net

The U-Net CNN architecture is a fully convolutional network (FCN) that has two main components: a contraction path, also called an encoder, which captures the image information; and the expansion path, also called decoder, which uses the encoded information to create the segmentation output [21]. In our case, it outputs a binary mask indicating the lungs. Figure 4 presents the U-Net architecture.

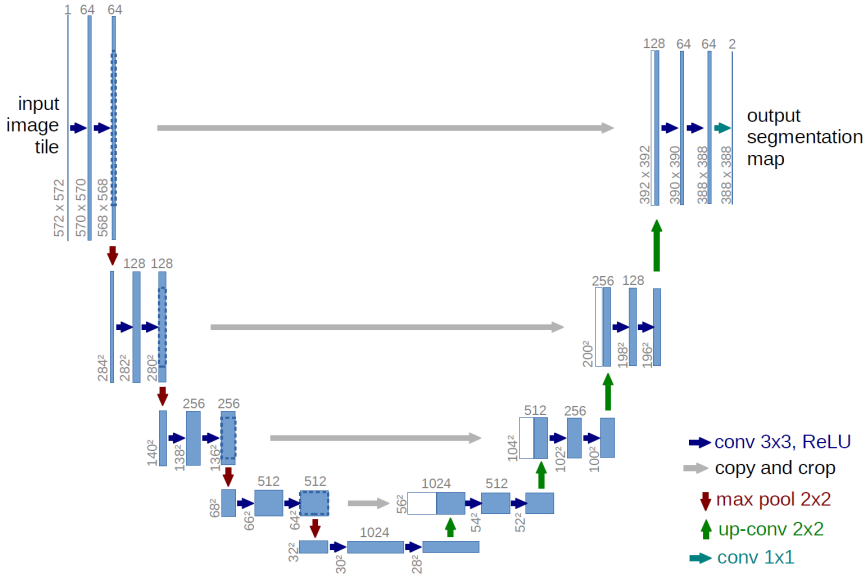



Figure 4: Original U-Net architecture [21]

#### 4.1.2 Additional CXR Masks

Deep learning techniques are very data-hungry, i.e., the more data we use for training, the better the model can learn any patterns. Hence, we also incorporated three additional CXR datasets that had manually created binary masks to increase our training sample size for segmentation only. These additional datasets were not used for COVID-19 identification. Their sole purpose was to improve the segmentation performance as a sort of transfer learning.

The first additional CXR dataset is Montgomery County X-ray Set, it contains a total of 138 posterior-anterior (PA) X-rays with manually created binary masks that indicate the lung area [23]. The second additional CXR dataset is Shenzhen Hospital X-ray Set, it contains a total of 662 X-rays [23]. A total of 566 masks were manually created by Jaeger et al. [24]. The last additional CXR dataset is Japanese Society of Radiological Technology (JSRT), it contains a total of 247 X-rays [25].

### 4.2 Classification (Phase 2)

We used only end-to-end deep learning approaches for classification without any handcrafted feature extraction algorithms. The reason behind this choice is to as much as possible preserve the original image to experiment with some prediction explanation approaches to show the impact of lung segmentation. This point will be further discussed in section 4.3.

[7]https://github.com/v7labs/COVID-19-xray-dataset

We chose a simple and straightforward approach with three of the most popular CNN architectures: VGG16, ResNet50V2 InceptionV3. For all, we used transfer learning by applying pre-trained weights from ImageNet provided by Keras. We performed the classification using full and segmented CXR images independently.

Furthermore, we also evaluated two specific scenarios to assess any bias in our proposed classification schema. First, we built a specific validation approach to assess the COVID-19 generalization from different sources, i.e., we want to answer the following question: is it possible to use COVID-19 CXR images from one database to identify COVID-19 in another different database? This scenario is our main contribution since it is the less database biased experiment in this paper.

Then, we also evaluated a database classification scenario, in which we used the database source as the final label, and used full and segmented CXR images to verify if lung segmentation reduces the database bias. We want to answer the following question: does lung segmentation reduces the underlying differences from different databases which might bias a COVID-19 classification model?

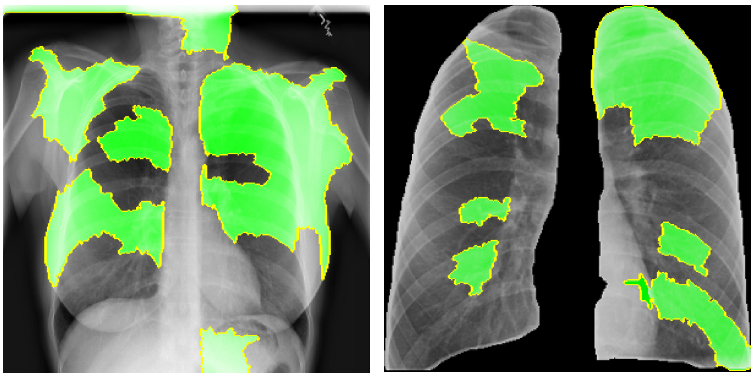
In the literature, many papers employ complex classification approaches. However, a complex model does not necessarily mean better performance whatsoever. Even very simple deep architectures tend to overfit very quickly [57]. There must be a solid argument to justify applying a complicated approach to a low sample size problem. Additionally, CXR images are not the gold standard for pneumonia diagnosis because it has low sensitivity [58, 6]. Thus, human performance in this problem is usually not very high [59]. That makes us wonder how realistic are some approaches presented in the literature, in which they achieve a very high classification accuracy.

### 4.3 XAI (Phase 3)

Depending on the perspective, most machine learning models can be seen as a black-box classifier, it receives input and somehow computes an output [41]. It might happen both with deep and shallow learning, with some exceptions like decision trees. Even though we can measure our model's performance using a set of metrics, it is nearly impossible to make sure that the model focuses on the correct portion of the test image for prediction.

Specifically, in our use case, we want the model to focus exclusively on the lung area and not somewhere else. If the model uses information from other regions, even if very high accuracy is achieved, there can be some limitations to its application, since it is not learning to identify COVID-19 but something else.

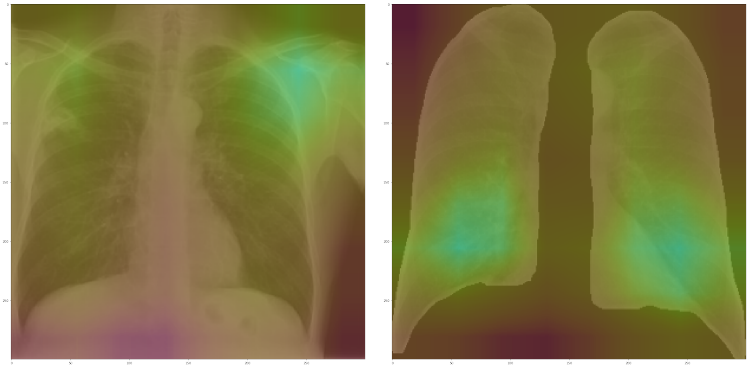
Here, we aim to demonstrate that by using segmented images, the model prediction uses primarily the lung area, which is not often the case when we use full CXR images. To do so, we applied two XAI approaches: LIME and Grad-CAM. Despite having the same main objective, they differ in how they find the important regions. Figures 5 and 6 shows examples of important regions highlighted by LIME and Grad-CAM, respectively. In section 6, we will show that models trained using segmented lungs focus primarily on the lung area, while models trained using full CXR images frequently focus elsewhere.



(a) Full CXR image. (b) Segmented CXR image.

Figure 5: LIME example.

The reason for not using handcrafted feature extraction algorithms here is that it is usually not straightforward to rebuild the reverse path, i.e., from prediction to the raw image. Sometimes, the handcrafted algorithm creates global features, eliminating the possibility of identifying the image regions that resulted in a specific feature.



(a) Full CXR image. (b) Segmented CXR image.

Figure 6: Grad-CAM example.

## 5 Experimental Setup

This section presents details about the proposed database, algorithms for segmentation and classification, parameters, and metrics used in this paper.

### 5.1 Lung Segmentation Database

Table 1 presents the main characteristics of the database used to perform experimentation on lung segmentation. It comprises 1,645 CXR images, with a 95/5 percentage train/test split. We also create a third set for training evaluation, called validation set, containing 5 percent of the training data.

Table 1: Lung segmentation database.

| *Characteristic* | *Samples* |
|---|---|
| Train | 1,483 |
| Validation | 79 |
| Test | 83 |
| Total | 1,645 |

As mentioned in Section 4.1.2, we combined multiple datasets that already had manually created lung masks to improve the segmentation model performance. Table 2 presents the samples distribution for each source.

Table 2: Lung segmentation database composition.

| *Source* | *Samples* |
|---|---|
| Cohen v7labs[8] | 489 |
| Montgomery | 138 |
| Shenzhen | 566 |
| JSRT | 247 |
| Manually created | 205 |

### 5.2 COVID-19 Database (RYDLS-20-v2)

Table 3 presents the main characteristics of the proposed database, which was named RYDLS-20-v2. The database comprises 2,678 CXR images, with an 80/20 percentage train/test split following a holdout validation split.

Considering the current effervescence of research using CXR image databases, the need for researchers to be frequently reviewing their protocols is somehow expected. In [59], the authors describe some issues regarding a previous version of that paper published by themselves. In that case, the authors figured out that by using a random cross-validation protocol on investigations using CXR images, there is a risk to accidentally place different images from the same

patient on different folds. This situation could lead to data leakage, which is undesirable. Aiming to prevent this situation, and also considering that there is not a consolidated database to be considered as a benchmark in this research topic, we decided not to use the cross-validation protocol in this work, even though we know that it is, in general, a recommendable practice in pattern classification modeling.

Therefore, we performed the split considering some crucial aspects: i) multiple CXR images from the same patient are always kept in the same fold; ii) images from the same source are evenly distributed in the train and test split; and iii) each class is balanced as much as possible while complying with the two previous restrictions. We also created a third set for training evaluation, called validation set, containing 20 percent of the training data randomly.

In this context, given the considerations mentioned above, simple random cross-validation would not suffice since it might not correctly separate the train and test split to avoid data leakage, and it could reduce robustness instead of increasing it. In this context, the holdout validation is a more comfortable option to ensure a fair and proper separation of train and test data. The test set was created to represent an independent test set in which we can validate our classification schema's generalization performance and evaluate the segmentation impact in a less biased database.

Table 3: RYDLS-20-v2 main characteristics.

| *Class* | *Train* | *Validation* | *Test* |
|---|---|---|---|
| Lung opacity | 739 | 189 | 231 |
| COVID-19 | 315 | 93 | 95 |
| Normal | 673 | 150 | 193 |
| **Total** | **1727** | **432** | **519** |

We built our database by further expanding our previous work RYDLS-20 [9] and adopting some guidelines and images provided by the COVIDx dataset [11]. Moreover, we set up the problem with three classes: lung opacity (pneumonia), COVID-19, and normal. We also experimented with expanding the number of classes to represent a more specific pathogen, such as bacteria, fungi, viruses, COVID-19, and normal. However, in all cases, the trained models did not differentiate between bacteria, fungi, and viruses very well, possibly due to the reduced sample size. Thus, we decided to take a more general approach to create a more reliable classification schema while retaining the focus on developing a more realistic approach.

The CXR images were obtained from eight different sources. Table 4 presents the samples distribution for each source.

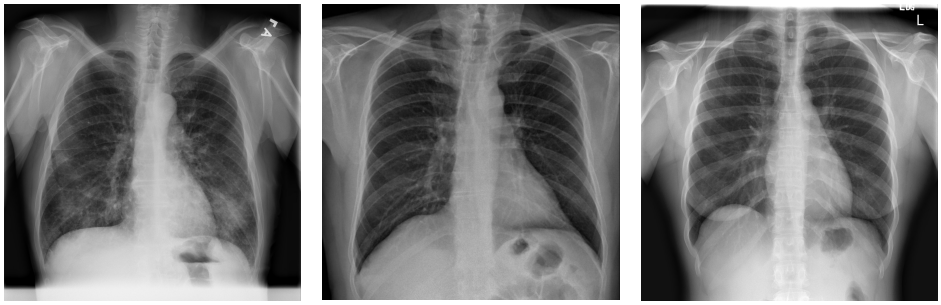
Table 4: Sources used in RYDLS-20-v2 database.

| *Source* | *Lung opacity* | *COVID-19* | *Normal* |
|---|---|---|---|
| Dr. Joseph Cohen GitHub Repository [49] | 140 | 418 | 16 |
| Kaggle RSNA Pneumonia Detection Challenge[9] | 1000 | - | 1000 |
| Actualmed COVID-19 Chest X-ray Dataset Initiative[10] | - | 51 | - |
| Figure 1 COVID-19 Chest X-ray Dataset Initiative[11] | - | 34 | - |
| Radiopedia encyclopedia[12] | 7 | - | - |
| Euroad[13] | 1 | - | - |
| Hamimi's Dataset[60] | 7 | - | - |
| Bontrager and Lampignano's Dataset [61] | 4 | - | - |

We considered posteroanterior (PA) and anteroposterior (AP) projections with the patient erect, sitting, or supine on the bed. We disregarded CXR with a lateral view because they are usually used only to complement a PA or AP view [62]. Additionally, we also considered CXR taken from portable machines, which usually happens when the patient cannot move (e.g., ICU admitted patients). This is an essential detail since there are differences between regular X-ray machines and portable X-ray machines regarding the image quality; we found most portable CXR images in the classes COVID-19 and lung opacity. We removed images with low resolution and overall low quality to avoid any issues when resizing the images.

Finally, we have no further details about the X-ray machines, protocols, hospitals, or operators, and these details impact the resulting CXR image. All CXR images are de-identified[14], and for some of them, there are demographic information available, such as age, gender, and comorbidities.

Figure 7 presents image examples for each class retrieved from the RYDLS-20-v2 database.



(a) Lung opacity. (b) COVID-19. (c) Normal.

Figure 7: RYDLS-20-v2 image samples.

As pointed out in the literature [12, 13, 14], it is critical to evaluate the database bias to ensure that our classification schema is classifying COVID-19 and not the database source. Besides, considering that we have three sources containing COVID-19 CXR images, we can also evaluate the generalization ability from one database to another.

### 5.3 COVID-19 Generalization

The COVID-19 generalization intents to demonstrate that our classification schema can identify COVID-19 in different CXR databases. To do so, we set up a binary problem with COVID-19 as the relevant class with a 2-fold validation using only segmented CXR images. The first fold contains all COVID-19 images from the Cohen database and a portion of the RSNA Kaggle database and the second fold contains the remaining RSNA Kaggle database and the other sources. Table 5 shows the samples distribution by source for this experiment. The primary purpose is to evaluate if the CXR images in the Cohen database allows the training of a non-random CNN classifier for the remaining COVID-19 source images and vice versa.

Table 5: COVID-19 generalization database composition.

| Source | Fold 1 | | Fold 2 | |
|---|---|---|---|---|
| | *Negative* | *COVID-19* | *Negative* | *COVID-19* |
| Dr. Joseph Cohen GitHub Repository | 156 | 418 | - | - |
| Kaggle RSNA Pneumonia Detection Challenge | 1000 | - | 1000 | - |
| Actualmed COVID-19 Chest X-ray Dataset Initiative | - | - | - | 51 |
| Figure 1 COVID-19 Chest X-ray Dataset Initiative | - | - | - | 34 |
| Radiopedia encyclopedia | - | - | 7 | - |
| Euroad | - | - | 1 | - |
| Hamimi's Dataset | - | - | 7 | - |
| Bontrager and Lampignano's Dataset | - | - | 4 | - |
| **Total** | 1156 | 418 | 1019 | 85 |

We must highlight that, despite this scenario being our least biased experiment, Kaggle RSNA is used in both folds, so it is not completely bias-free.

[14] Aiming at attending to data privacy policies.

### 5.4 Database Bias

Moreover, we also evaluated a dataset classification to assess if a CNN can identify the CXR image source using segmented and full CXR images. To do so, we set up a multi-class classification problem with three classes, one for each relevant image source: `Cohen`, `RSNA`, and `Other` (the remaining images from other sources combined). The database comprises 2,678 CXR images, with an 80/20 percentage of train/test split following a random holdout validation split. For training evaluation, we also created a validation set containing 20 percent of the training data randomly. The number of samples distributed among these sets for each data source is presented in Table 6.

Table 6: Database bias evaluation composition.

| *Class* | *Train* | *Validation* | *Test* |
|---|---|---|---|
| Cohen | 364 | 89 | 121 |
| RSNA | 1288 | 326 | 386 |
| Other | 61 | 14 | 29 |
| **Total** | **1713** | **429** | **536** |

The rationale is to assess if the database bias is reduced when we use segmented CXR images instead of full CXR images. Such evaluation is of great importance to ensure that the model classifies the relevant class, in this case, COVID-19, and not the image source.

### 5.5 Algorithms, Parameters and Metrics

This section presents details regarding the algorithms, parameters, and metrics used throughout this paper.

#### 5.5.1 U-Net

As previously mentioned, for the lung segmentation, we used the U-Net CNN architecture with some small differences: we included dropout and batch normalization layers in each contracting and expanding block. These additions aim to improve training time and reduce overfitting. Figure 8 presents our adapted U-Net architecture.

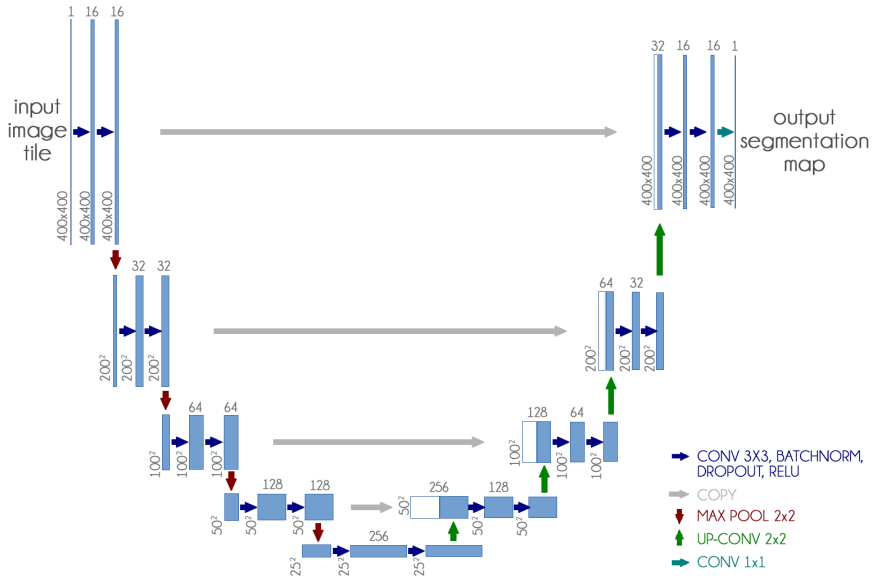



Figure 8: Custom U-Net architecture

Furthermore, since our dataset is not standardized, the first step was to resize all images to 400px $\times$ 400px. We chose that dimension because it presented a good balance between computational requirement and classification performance. We also experimented with smaller and larger dimensions with no significant improvement.

We achieve a much better result without using transfer learning and training the network weights from scratch for this model.

Table 7 reports the parameters used in U-Net training. We also used a Keras callback to reduce the learning rate by half once learning stagnates for three consecutive epochs.

Table 7: U-Net parameters.

| *Parameter* | *Value* |
|---|---|
| Epochs | 100 |
| Batch size | 16 |
| Learning rate | 0.001 |

After the segmentation, we applied a morphological opening with 5 pixels to remove small brights spots, which usually happened outside the lung region. We also applied a morphological dilation with 5 pixels to increase and smooth the predicted mask boundary. Finally, we also cropped all images to keep only the ROI indicated by the mask. After crop the images were also resized to 300px $\times$ 300px. Figure 2 shows an example of this process.

#### 5.5.2 Classification

We have used three well-known CNN architectures from the literature to perform the multi-class classification task: VGG16, ResNet50V2, and InceptionV3. In all cases, we applied transfer learning by loading pre-trained weights from ImageNet only for the convolutional layers [63]. We then added three fully-connected (FC) layers together, followed by dropout and batch normalization layers containing 1024, 1024, and 512 units.

For the training, we followed the typical workflow used in the literature [63]: i) in the first step, which we called warm-up, we froze all convolutional layers and trained the FC layers; ii) after that, in fine-tuning, the convolutional were unfroze all layers and training the entire network with a lower learning rate.

Table 8 reports the parameters used in the CNN training. We also used a Keras callback to reduce the learning rate by half once learning stagnates for three consecutive epochs.

Table 8: CNN parameters.

| *Parameter* | *Value* |
|---|---|
| Warm-up epochs | 50 |
| Fine-tuning epochs | 100 |
| Batch size | 40 |
| Warm-up learning rate | 0.001 |
| Fine-tuning learning rate | 0.0001 |

#### 5.5.3 Data Augmentation

We extensively used data augmentation during training in segmentation and classification to virtually increase our training sample size [64]. Table 9 presents the transformations used during training along with their parameters. The probability of applying each transformation was kept at the default value of 50%. We used the library *albumentations*[15] to perform all transformations [65]. Figure 9 displays some examples of the transformations applied.

#### 5.5.4 Evaluation Metrics

In this paper, there are two experiments: lung segmentation and classification. For each, we chose different metrics to analyze the performance of the experimental results.

For the lung segmentation task, we used two well-known metrics [66]: Jaccard distance and Dice coefficient. The Jaccard distance measures the dissimilarity between the ground-truth mask and the predicted mask; thus, a lower value is better. It is calculated by subtracting the Jaccard index from one. The Jaccard index, also known as Intersection over Union (IoU), measures the similarity between the ground-truth mask and the predicted mask, and it is defined as the intersection divided by the union of the two masks. We used the Jaccard distance as the loss function during training. The Dice coefficient is defined as $2 \times$ intersection divided by the union. It is somehow similar to the concept of F1-Score. Figure 10 shows a visual representation of the metrics used for segmentation using Venn diagrams. Both metrics work very similarly. The significant difference is that the Jaccard index and distance penalizes wrong predictions more than the Dice coefficient.

[15] https://github.com/albumentations-team/albumentations

Table 9: Data augmentation parameters.

| *Transformation* | *Segmentation* | *Classification* |
|---|---|---|
| Horizontal flip | – | – |
| Shift scale rotate | Shift limit = 0.0625<br>Scale limit = 0.1<br>Rotate limit = 45 | Shift limit = 0.05<br>Scale limit = 0.05<br>Rotate limit = 15 |
| Elastic transform | Alpha = 1<br>Sigma = 50<br>Alpha affine = 50 | Alpha = 1<br>Sigma = 20<br>Alpha affine = 20 |
| Random brightness | Limit = 0.2 | Limit = 0.2 |
| Random contrast | Limit = 0.2 | Limit = 0.2 |
| Random gamma | Limit = (80, 120) | Limit = (80, 120) |

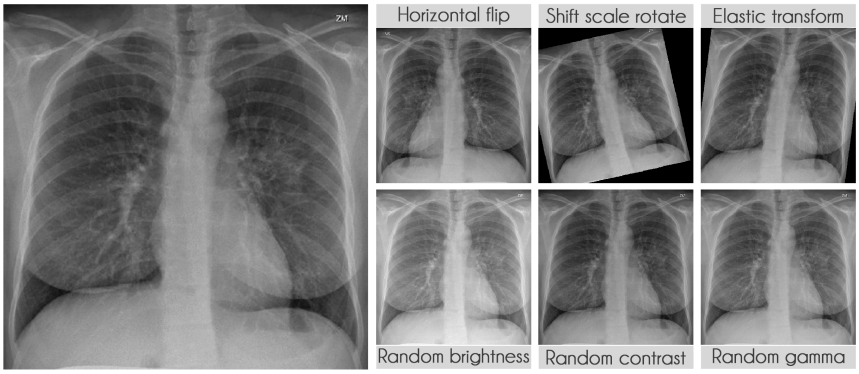
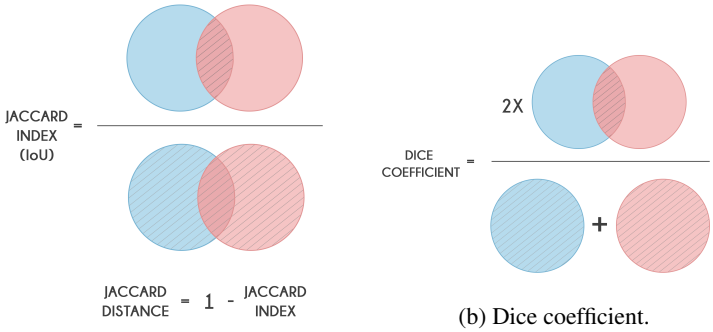



Figure 9: Data augmentation examples.




(a) Jaccard metrics.

(b) Dice coefficient.

Figure 10: Segmentation metrics.

For the classification task, we used precision, recall, and F1-Score. Precision is the proportion of predicted positives that are actually positive. Recall is the proportion of positive correctly classified. The F1-Score is defined as the harmonic average between precision and recall. The mathematical definition of the metrics are

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3}$$

Where $TP$ is the true positive rate, $FP$ is the false positive rate, and $FN$ is the false-negative rate. Additionally, we used the macro-averaged evaluation to summarize the classification performance for all classes. The macro-averaged approach averages the F1-Score per class [67].

#### 5.5.5 XAI

For each image in the test set, we used LIME and Grad-CAM to find the most important regions used for the predicted class, i.e., regions that support the given prediction. We then summarized all those regions in a heatmap to show the most common regions that the model uses for prediction. Thus, we have one heatmap per classifier per class per XAI approach.

Table 10 presents the parameters used for the LIME explanation. Grad-CAM has a single configurable parameter, which is the convolutional layer to be used, and, in our case, we always used the last convolutional layer before the FC layers.

Table 10: LIME parameters.

| *Parameter* | *Value* |
|---|---|
| Superpixels identification | Quickshift segmentation |
| Quickshift kernel size | 4 |
| Distance metric | Cosine |
| Number of samples per image | 1000 |
| Number of superpixels in explanation per image | 5 |
| Filter only positive superpixels | True |

## 6 Experimental Results

This section presents an overview of our experimental findings and a preliminary analysis of each contribution individually.

### 6.1 Lung Segmentation Results

Table 11 shows the overall U-Net segmentation performance for the test set for each source we used to compose the lung segmentation database considering the Jaccard distance and the Dice coefficient metrics.

Table 11: Lung segmentation results.

| *Database* | *Jaccard distance* | *Dice coefficient* |
|---|---|---|
| Cohen v7labs | $0.041 \pm 0.027$ | $0.979 \pm 0.014$ |
| Montgomery | $0.019 \pm 0.007$ | $0.991 \pm 0.003$ |
| Shenzhen | $0.017 \pm 0.008$ | $0.991 \pm 0.004$ |
| JSRT | $0.018 \pm 0.011$ | $0.991 \pm 0.006$ |
| Manually created masks | $0.071 \pm 0.021$ | $0.964 \pm 0.011$ |
| Test set | $0.035 \pm 0.027$ | $0.982 \pm 0.014$ |

As we expected, our manually created masks had poor performance compared to the other sources' results, mainly because we are not professional radiologists and our created masks were not as good as the other sources. Following

that, the Cohen v7labs set also presented a somewhat lower performance. Our manual inspection showed that the model did not include the heart region for CXR images, while this database always included the heart, hence the difference. The performance of the remaining databases is outstanding.

### 6.2 Multi-class Classification

Table 12 presents F1-Score results for our multi-class scenario. The models using non-segmented CXR images presented better results than the models that used segmented images when we consider raw performance for COVID-19 and lung opacity. Both settings were on par in the normal class.

Table 12: F1-Score results.

| *Class* | *COVID-19* | *Lung opacity* | *Normal* | *Macro-avg* |
|---|---|---|---|---|
| Segmented - VGG16 | 0.83 | 0.88 | 0.9 | 0.87 |
| Segmented - ResNet50V2 | 0.78 | 0.87 | 0.91 | 0.85 |
| Segmented - InceptionV3 | 0.83 | 0.89 | 0.92 | 0.88 |
| Non-segmented - VGG16 | 0.94 | 0.91 | 0.91 | 0.92 |
| Non-segmented - ResNet50V2 | 0.91 | 0.9 | 0.92 | 0.91 |
| Non-segmented - InceptionV3 | 0.86 | 0.9 | 0.91 | 0.9 |

In all cases, the models using segmented images performed worse, considering the selected metric. That result alone might discourage the usage of segmentation in practice. However, in Section 6.5, we will show that it is still worth to take into account the segmentation strategy. Even though the use of segmentation does not lead to improvements in the F1-Score rates, the resulting models may present a more realistic performance.

### 6.3 COVID-19 Generalization

Table 13 shows the F1-Score results for the COVID-19 generalization. The classification was set up as a binary problem with COVID-19 as the positive class in this problem. The folds were separated in a way that the COVID-19 CXR images from the Cohen database would not be in the same fold of COVID-19 CXR images from the two other databases that contain COVID-19 cases (Actualmed and Figure1 GitHub repositories). The results are auspicious and indeed show that classification, in this case, is far from random. We achieved an F1-Score of 0.77 and 0.7 in the first and second folds, respectively. The lower performance in the second fold was somewhat expected since it contains few COVID-19 examples for training. Figure 11 presents the ROC curve for this scenario.

Table 13: F1-Score COVID-19 generalization results.

| *Model* | *Fold 1* | *Fold 2* | *Macro-avg* |
|---|---|---|---|
| VGG16 | 0.76 | 0.65 | 0.71 |
| ResNet50V2 | 0.77 | 0.68 | 0.73 |
| InceptionV3 | 0.77 | 0.70 | 0.74 |

### 6.4 Database Bias

Table 14 shows the F1-Score results for the database bias evaluation. In this problem, the classification was set up as a multi-class problem with database source as the corresponding label for full and segmented CXR images. The results show that overall the lung segmentation reduces the differences between databases. However, even after segmentation, it is possible to identify the source with fair confidence. Such a result may be because the majority of some classes are extracted from the same databases. For instance, most COVID-19 CXR images are from Cohen, and most normal CXR images are from RSNA. Hence in this situation, it is hard to isolate and measure both effects. Furthermore, the class Other contains six different sources, so it is unfair to compare it to Cohen or RSNA. Thus the macro-averaged F1-Score presented does not take it into account. In conclusion, this highlights the need for a bigger and more comprehensive COVID-19 CXR database.

### 6.5 XAI Results

Figures 12 and 13 present the LIME and Grad-CAM heatmaps for our multi-class scenario. We can notice that the models created using segmented CXR images focused primarily in the lung area. The lung shape is discernible in all

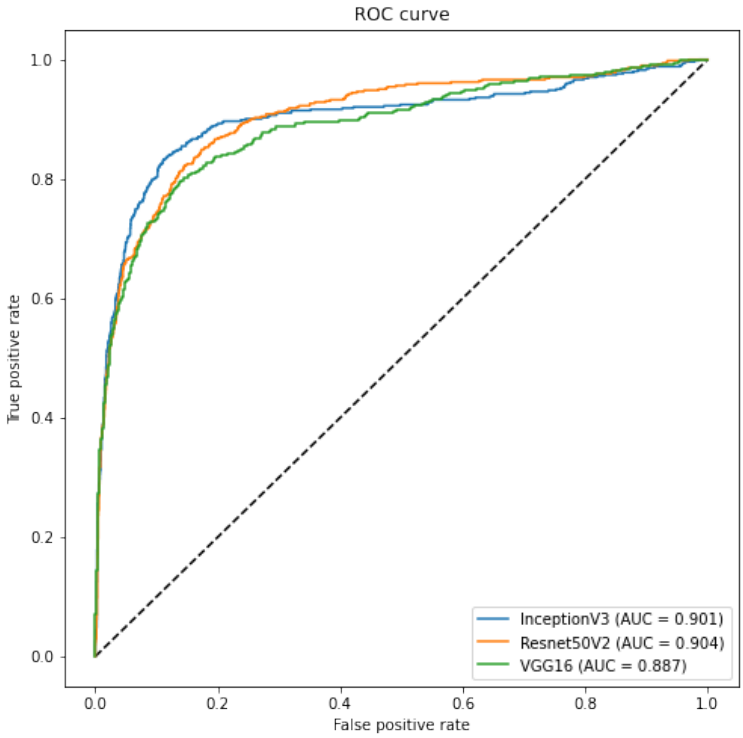



Figure 11: COVID-19 Generalization ROC Curve.

Table 14: F1-Score database bias results.

| *Scenario* | *Cohen* | *RSNA* | *Other* | *Macro-avg** |
|---|---|---|---|---|
| Segmented - VGG16 | 0.65 | 0.91 | 0 | 0.78 |
| Segmented - ResNet50V2 | 0.62 | 0.9 | 0.07 | 0.76 |
| Segmented - InceptionV3 | 0.61 | 0.89 | 0.24 | 0.75 |
| Non-segmented - VGG16 | 0.89 | 0.98 | 0.61 | 0.93 |
| Non-segmented - ResNet50V2 | 0.85 | 0.97 | 0 | 0.91 |
| Non-segmented - InceptionV3 | 0.88 | 0.98 | 0.53 | 0.93 |

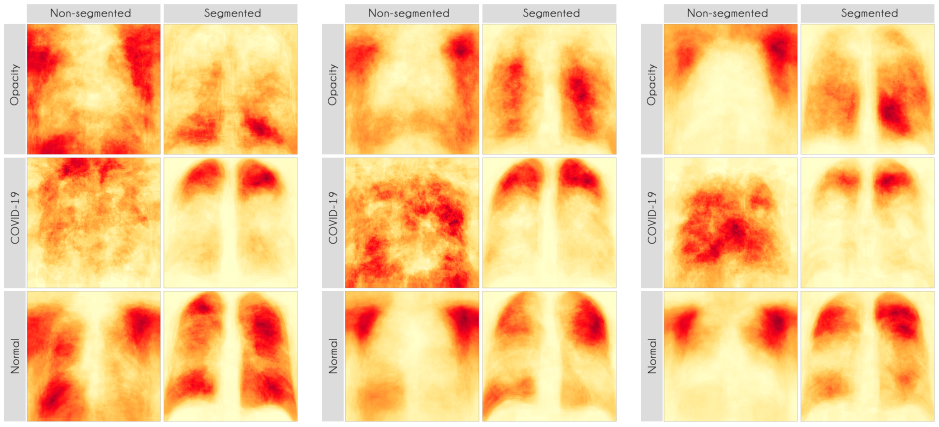
*Macro-averaged F1-Score for Cohen and RSNA.

heatmaps. The only small exception is the VGG16 Lung Opacity class. Despite having the visible lung shape, it also focused a lot in other regions. In contrast, the models that used full CXR images are more chaotic. We can see, for instance, that for both InceptionV3 and VGG16, the Lung Opacity and Normal class heatmaps almost did not focus on the lung area at all.

Even though the models that used full CXR images performed better, considering the F1-Score, they used information outside the lung area to predict the output class. Thus, they did not necessarily learn to identify lung opacity or COVID-19, but something else. Hence, we can say that even though they perform better, considering the classification metric, they are worse and not reliable for real-world applications.

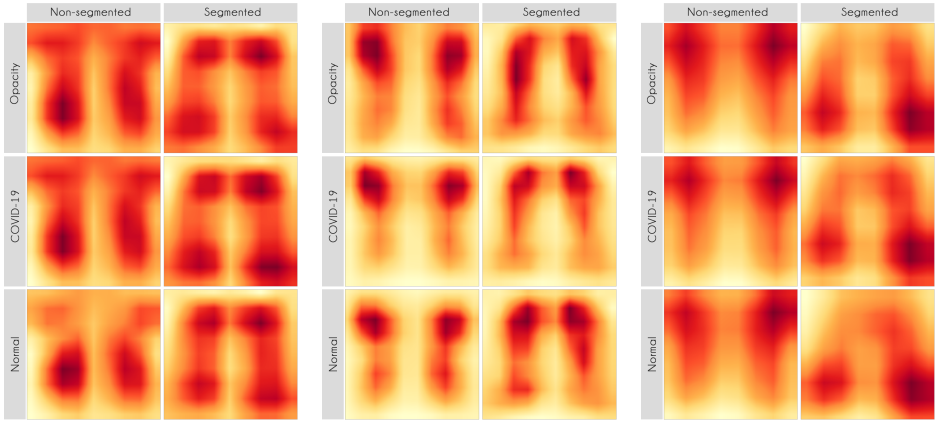
## 7 Discussions

This section discusses the importance and significance of the results obtained. Given that we have multiple experiments, we decided to create subsections to drive the discussion better.




(a) VGG16. (b) ResNet50V2. (c) InceptionV3.

Figure 12: LIME heatmaps.




(a) VGG16. (b) ResNet50V2. (c) InceptionV3.

Figure 13: Grad-CAM heatmaps.

### 7.1 Multi-class Classification

A Wilcoxon signed-rank test indicated that the models using segmented CXR images have a significantly lower F1-Score than the models using non-segmented CXR images ($p = 0.019$). Additionally, a Bayesian t-test also indicated that using segmented CXR images reduces the F1-Score with a Bayes Factor of 2.1. The Bayesian framework for hypothesis testing is very robust even for a low sample size [68]. Figure 14 presents a visual representation of our classification results stratified by lung segmentation with a boxplot.

In general, models using full CXR images performed significantly better, which is an exciting result since we expected otherwise. This result was the main reason we decided to apply XAI techniques to explain individual predictions. Our rationale is that a CXR image contains a lot of noise and background data, which might trick the classification model into focusing on the wrong portions of the image during training. Figure 15 presents some examples of the Grad-CAM explanation showing that the model is actively using burned in annotations for the prediction. The LIME heatmaps presented in Figure 12 show that exactly behavior for the classes Lung opacity and Normal in the non-segmented models, i.e., the model learned to identify the annotations and not lung opacities. The Grad-CAM heatmaps in Figure 13 also show the focus on the annotations for all classes in the non-segmented models.

The most affected class by lung segmentation is the COVID-19, followed by Lung opacity. The Normal class had a minimal impact. The best F1-Scores for COVID-19 and Lung opacity using full CXR images are 0.94 and 0.91, respectively, and after the segmentation, they are 0.83 and 0.89, respectively. We conjecture that such impact comes from the fact that many CXR images are from patients with severe clinical conditions who cannot walk or stand. Thus

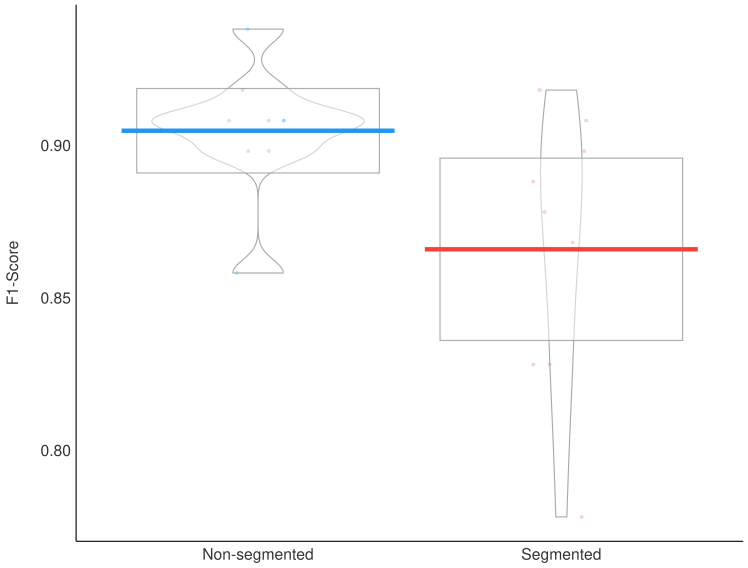
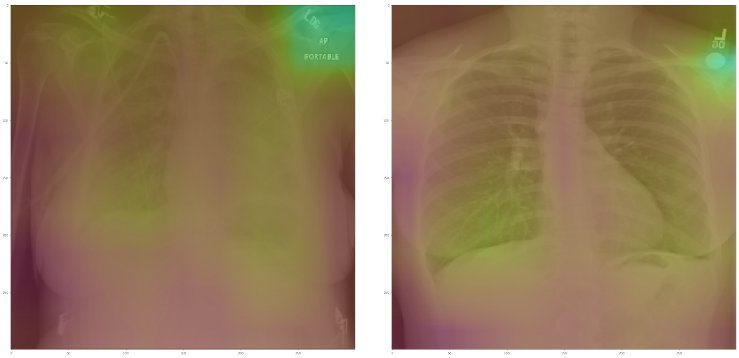



Figure 14: F1-Score results boxplot stratified by segmentation.



(a) Example 1. (b) Example 2.

Figure 15: Grad-CAM showing a large gradient on CXR annotations.

the medical practitioners must use a portable X-ray machine that produces images with the "AP Portable" annotation. That impact also means that the classification models had trouble identifying COVID-19.

Considering specifically the models using segmented CXR images, InceptionV3 performed better in all classes. Figure 16 provides a visual representation of the F1-Score achieved in the experimental results stratified by the model used and lung segmentation. Figure 17 shows the confusion matrix for the InceptionV3 using segmented CXR images. Overall the classifier presented a remarkable performance in all labels. The largest misclassification happened with the class Lung opacity being predicted as Normal, followed by the class COVID-19 being predicted as Lung opacity. However, there are reasonable explanations for both: i) Most examples from the classes Lung opacity and Normal came from the RSNA database; thus, we believe that the data source biased the classification marginally; ii) pneumonia caused by COVID-19 could have been confused with pneumonia caused by another pathogen. A solution for both issues would be to increase the number of images in the database, including more data sources.

### 7.2 XAI

In this paper, we applied two XAI techniques: LIME and Grad-CAM. The reason for applying both is to evaluate the classification models thoroughly since they work differently. They have some significant differences and highlights: i) LIME is model-agnostic, and Grad-CAM is model-specific; ii) in LIME, the granularity of important regions is correlated to the granularity of the superpixel identification algorithm; iii) Grad-CAM produces a very smoothed output

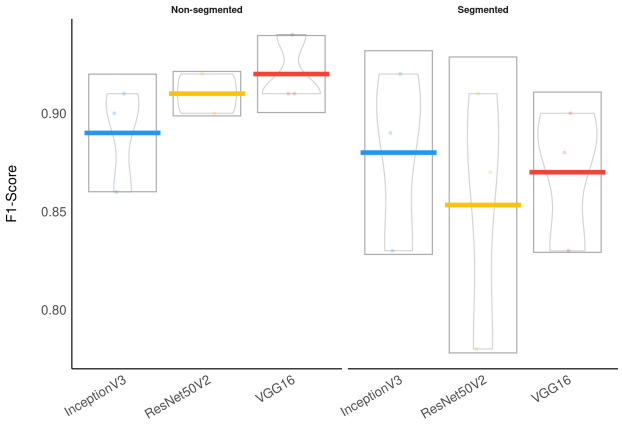



Figure 16: F1-Score results boxplot stratified by segmentation and model.

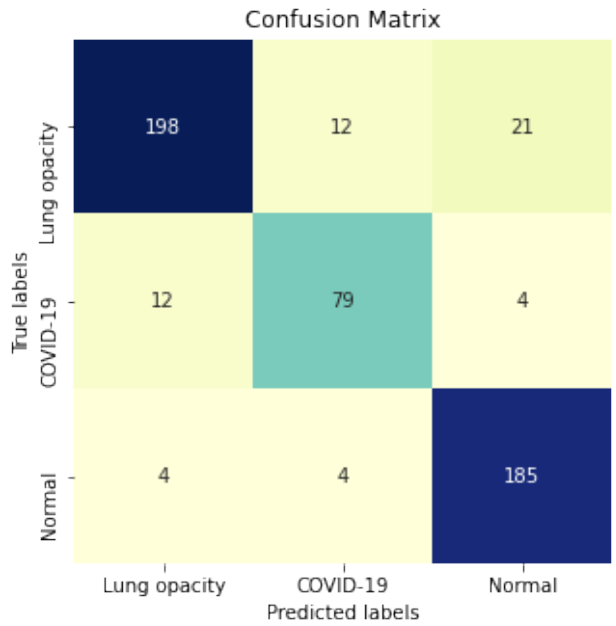



Figure 17: Segmented InceptionV3 Confusion Matrix.

because the dimension of the last convolution layer is much smaller than the dimension of the original input. Keep in mind that such techniques are not definitive. They can complement and corroborate with each other. Thus, we can increase the model reliability in a real-world context by using a more comprehensive approach.

Our XAI approach is novel in the sense that we explored a more general explanation instead of focusing on single examples. In the literature, there are many papers exploring LIME and Grad-CAM for a couple of handpicked examples. The main problem with such approaches is that the examples might have been eventually chosen to reach a specific result. In this paper, we applied the XAI techniques to each image in the test set individually and created a heatmap aggregating all individual results to represent a broader context, which indicates which portions of the CXR image the models have focused on for prediction. Figures 12 and 13 demonstrate that the models using full CXR images are misleading because they focus a lot on the left and right uppermost regions, which is usually the location of burned-in annotations.

### 7.3 COVID-19 Generalization and Database Bias

The multi-class scenario is fascinating to visualize the behavior of individual models. However, given the strong database bias present in this context, even after lung segmentation, the multi-class results are not entirely reliable.

In order to evaluate such bias and provide a more realistic result, we crafted two specific scenarios to ensure that our classification model is not classifying the database source. First, as we have multiple sources of COVID-19 CXR images, we verified if it was possible to use CXR images from one database to train a model to recognize COVID-19 CXR in the other databases. We achieved a macro-averaged F1-Score of 0.74 using InceptionV3 and an area under the ROC curve of 0.9 using InceptionV3 and ResNet50V2. The F1-Score was lower than in our multi-class scenario. However, this corroborates that it is possible to identify COVID-19 cases across databases, i.e., our classification model is indeed identifying COVID-19 and not the database source. Such a scenario constitutes our main result and contribution, since it represents a less biased and more realistic performance, given the hurdles that still exist with COVID-19 CXR databases.

Second, as discussed in the work of [13], there is a strong bias towards the database source in this context. In our evaluation, we found out that lung segmentation consistently reduces the ability to differentiate the sources. We achieved a database classification F1-Score of 0.93 and 0.78 for full and segmented CXR images, respectively. A Wilcoxon signed-rank test and a Bayesian t-test indicated that segmentation reduces the macro-averaged F1-Score with statistical significance ($p = 0.024$ and a Bayes Factor of 4.6). Despite that, even after segmentation, there is a strong bias towards the RSNA Kaggle database, considering specifically this class, we achieved an F1-Score of 0.91. In summary, the usage of lung segmentation is outstanding in reducing the database bias in our context. However, it does remedy the issue entirely.

### 7.4 Concluding Remarks

In a real-world application, especially in medical practice, we must be cautious and thorough when designing systems aimed at diagnostic support because they directly affect people's lives. A misdiagnosis can have severe consequences for the health and further treatment of a patient. Furthermore, in the COVID-19 pandemic, such consequences can also affect other people since it is a highly infectious disease. Even though the current pandemic attracted much attention from the research community in general, few works focused on a more critical evaluation of the solutions proposed.

Ultimately, we demonstrated that lung segmentation is essential for COVID-19 identification in CXR images through a comprehensive and straightforward application of deep models coupled with XAI techniques. In fact, in our previous work [9], we have addressed the task of pneumonia identification as a whole, stating that maybe the patterns of the injuries caused by the different pathogens (virus, bacteria, and fungus) are different, so we were able to classify the CXR images with machine learning techniques. Even though the experimental results of that work have shown that it may be possible, it is challenging to be sure that other patterns did not bias the results in the images that were not related to the lungs.

Furthermore, as previously noted, we still believe that even after lung segmentation, the database bias still marginally influenced the classification model. Thus, more aspects regarding the CXR images and the classification model must be further evaluated to design a proper COVID-19 diagnosis system using CXR images.

## 8 Conclusion

The application of pattern recognition techniques has proven to be very useful in many situations in the real world. Explicitly considering the COVID-19 pandemic, machine learning can be used to help diagnose this disease in the

population. Several papers propose using machine learning methods to identify pneumonia and COVID-19 in CXR images with encouraging results in this setting. However, very few proposed to use lung segmentation to avoid any data leak or overfitting, and only focused on the classification metric itself.

Considering a real-world application, segmentation is a crucial step since it removes background information, reduces the chance of data leak, and forces the model to focus only on important image areas. Furthermore, the segmentation might not improve the classification performance, but since it forces the model to use only the lung area information, it increases the model reliability and quality.

In this paper, our objective was to demonstrate that the lung segmentation should be a mandatory step before attempting any diagnosis using CXR images. To achieve that, we first improved our previous proposed dataset, named RYDLS-20-v2, by increasing the number of sources and images. We then trained three popular CNN architectures: VGG16, ResNet50V2, and InceptionV3, using full and segmented CXR images. Finally, we used XAI strategies (i.e., LIME and Grad-CAM) to find important regions used by the models to perform the predictions.

Our lung segmentation model presented robust performance considering two factors: i) we did not aim to surpass the state-of-the-art performance of lung segmentation in CXR images; instead, we focused on creating a general segmentation model capable of producing binary lung masks for CXR images in our COVID-19 database; ii) the lung segmentation database was composed of multiple sources, some masks were even manually created. Nevertheless, our approach was on par with current state-of-the-art lung segmentation in CXR images [69, 70, 71].

Furthermore, we applied LIME and Grad-CAM to demonstrate that using segmented CXR images, the models focused primarily on information in the lung area to classify the CXR images. Thus, despite lowering the F1-Score, segmentation improves the prediction quality as it forces the model to use only relevant information.

We do not want to claim state-of-the-art classification results at this time for a couple of reasons: i) there are some initiatives to build a comprehensive COVID-19 CXR database; however, we still do not have a reliable database that can be used as a definitive benchmark; ii) in clinical practice, a small difference in the classification performance is hardly noticeable, and the model reliability and quality are more important than the classification metric [72]; and, iii) the CXR is not the gold standard for diagnosis, even experienced medical practitioners sometimes face doubts when examining a CXR image [6]; thus we should be very cautious at papers claiming very high classification performance when the human performance is much lower.

Despite many initiatives to build a comprehensive COVID-19 database, there is no definitive COVID-19 benchmark database. Thus, it is unfair to make direct comparisons of identification rates from different works, as they usually use different databases under different circumstances. Nevertheless, to the best of our knowledge, we achieved the best identification rate of COVID-19 among other types of pneumonia using segmented CXR images in a less biased configuration. Additionally, we must highlight our novel approach to demonstrate the importance of lung segmentation in CXR images classification.

As future work, we aim to keep improving our database to increase our classification performance and provide more robust estimates by using more CNN architectures for segmentation and classification. Furthermore, we want to apply more sophisticated segmentation techniques to isolate specific lung opacities caused by COVID-19. Likewise, we also want to explore more approaches to evaluate the model predictions, such as SHAP [73].

## References

[1] World Health Organization WHO. Coronavirus disease (COVID-19): situation report, 209. 2020.

[2] Matthew Zirui Tay, Chek Meng Poh, Laurent Rénia, Paul A MacAry, and Lisa FP Ng. The trinity of COVID-19: immunity, inflammation and intervention. *Nature Reviews Immunology*, pages 1–12, 2020.

[3] Giacomo Grasselli, Antonio Pesenti, and Maurizio Cecconi. Critical care utilization for the COVID-19 outbreak in lombardy, italy: early experience and forecast during an emergency response. *Journal of the American Medical Association*, 323(16):1545–1546, 2020.

[4] Stephen A Lauer, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 2020.

[5] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B Munroe, Bina Joe, and Xi Cheng. Artificial intelligence and machine learning to fight COVID-19, 2020.

[6] Wesley H Self, D Mark Courtney, Candace D McNaughton, Richard G Wunderink, and Jeffrey A Kline. High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ed patients: implications for diagnosing pneumonia. *The American Journal of Emergency Medicine*, 31(2):401–405, 2013.

[7] Hayit Greenspan, Raúl San José Estépar, Wiro J Niessen, Eliot Siegel, and Mads Nielsen. Position paper on covid-19 imaging and ai: from the clinical needs and technological challenges to initial ai solutions at the lab and national level towards a new era for ai in healthcare. *Medical Image Analysis*, page 101800, 2020.

[8] Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. Deep-COVID: Predicting COVID-19 from chest x-ray images using deep transfer learning. *Medical Image Analysis*, 65:101794, 2020.

[9] Rodolfo M Pereira, Diego Bertolini, Lucas O Teixeira, Carlos N Silla Jr, and Yandre M G Costa. COVID-19 identification in chest x-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, 194:105532, 2020.

[10] Suat Toraman, Talha Burak Alakuş, and İbrahim Türkoğlu. Convolutional capsnet: A novel artificial neural network approach to detect covid-19 disease from x-ray images using capsule networks. *Chaos, Solitons & Fractals*, page 110122, 2020.

[11] Linda Wang and Alexander Wong. COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871. Unpublished results*, 2020.

[12] Beatriz Garcia Santa Cruz, Jan Sölter, Matias Nicolas Bossa, and Andreas Dominik Husch. On the composition and limitations of publicly available covid-19 x-ray imaging datasets. *arXiv preprint arXiv:2008.11572. Unpublished results*, 2020.

[13] Gianluca Maguolo and Loris Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *arXiv preprint arXiv:2004.12823. Unpublished results*, 2020.

[14] Enzo Tartaglione, Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, and Marco Grangetto. Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data. *arXiv preprint arXiv:2004.05405. Unpublished results*, 2020.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2015.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[18] RM Hopstaken, EE Stobberingh, JA Knottnerus, JWM Muris, P Nelemans, PELM Rinkens, and GJ Dinant. Clinical items not helpful in differentiating viral from bacterial lower respiratory tract infections in general practice. *Journal of Clinical Epidemiology*, 58(2):175–183, 2005.

[19] R Virkki, T Juven, H Rikalainen, E Svedström, J Mertsola, and O Ruuskanen. Differentiation of bacterial and viral pneumonia in children. *Thorax*, 57(5):438–441, 2002.

[20] Er Anjna Anjna and Rajandeep Kaur Er. Review of image segmentation technique. *International Journal of Advanced Research in Computer Science*, 8(4), 2017.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[22] Charu C Aggarwal. *Neural networks and deep learning; A textbook*. Springer, 2018.

[23] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475, 2014.

[24] Sergii Stirenko, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Yuri Gordienko, Peng Gang, and Wei Zeng. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In *Proceedings of the IEEE International Conference on Electronics and Nanotechnology*, pages 422–428. IEEE, 2018.

[25] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[26] Sema Candemir and Sameer Antani. A review on lung boundary detection in chest x-rays. *International Journal of Computer Assisted Radiology and Surgery*, 14(4):563–576, 2019.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[29] Jimmy Whitworth. COVID-19: a fast evolving pandemic. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 114(4):241, 2020.

[30] Marc Lipsitch, David L Swerdlow, and Lyn Finelli. Defining the epidemiology of covid-19—studies needed. *New England Journal of Medicine*, 382(13):1194–1196, 2020.

[31] Hussin A Rothan and Siddappa N Byrareddy. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of Autoimmunity*, page 102433, 2020.

[32] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675, 2020.

[33] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Chung, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology*, page 201160, 2020.

[34] Michael S Niederman, Lionel A Mandell, Antonio Anzueto, John B Bass, William A Broughton, G Douglas Campbell, Nathan Dean, Thomas File, Michael J Fine, Peter A Gross, et al. Guidelines for the management of adults with community-acquired pneumonia: diagnosis, assessment of severity, antimicrobial therapy, and prevention. *American Journal of Respiratory and Critical Care Medicine*, 163(7):1730–1754, 2001.

[35] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[36] Kevin Gurney. *An Introduction to Neural Networks*. CRC Press, 1997.

[37] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285. Springer, 1982.

[38] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, pages 319–345. Springer, 1999.

[39] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[40] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614. Unpublished results*, 2016.

[41] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 5686–5697, 2016.

[42] Fei Shan, Yaozong Gao, Jun Wang, Weiya Shi, Nannan Shi, Miaofei Han, Zhong Xue, and Yuxin Shi. Lung infection quantification of COVID-19 in ct images with deep learning. *arXiv preprint arXiv:2003.04655. Unpublished results*, 2020.

[43] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*, 2020.

[44] Xiaocong Chen, Lina Yao, and Yu Zhang. Residual attention u-net for automated multi-class segmentation of COVID-19 chest ct images. *arXiv preprint arXiv:2004.05645. Unpublished results*, 2020.

[45] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Chao-Wei Zhao, and Ming-Ming Cheng. Jcs: An explainable COVID-19 diagnosis system by joint classification and segmentation. *arXiv preprint arXiv:2004.07054. Unpublished results*, 2020.

[46] Yu Qiu, Yun Liu, and Jing Xu. Miniseg: An extremely minimum network for efficient COVID-19 segmentation. *arXiv preprint arXiv:2004.09750. Unpublished results*, 2020.

[47] Qingsen Yan, Bo Wang, Dong Gong, Chuan Luo, Wei Zhao, Jianhu Shen, Qinfeng Shi, Shuo Jin, Liang Zhang, and Zheng You. COVID-19 chest ct image segmentation–a deep convolutional neural network solution. *arXiv preprint arXiv:2004.10987. Unpublished results*, 2020.

[48] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic COVID-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 2020.

[49] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. COVID-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988. Unpublished results*, 2020.

[50] Md Karim, Till Döhmen, Dietrich Rebholz-Schuhmann, Stefan Decker, Michael Cochez, Oya Beyan, et al. Deepcovidexplainer: Explainable COVID-19 predictions based on chest x-ray images. *arXiv preprint arXiv:2004.04582. Unpublished results*, 2020.

[51] Jianpeng Zhang, Yutong Xie, Zhibin Liao, Guansong Pang, Johan Verjans, Wenxin Li, Zongji Sun, Jian He, Yi Li, Chunhua Shen, et al. Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection. *arXiv: 2003.12338. Unpublished results*, 2020.

[52] Narinder Singh Punn and Sonali Agarwal. Automated diagnosis of COVID-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *arXiv preprint arXiv:2004.11676. Unpublished results*, 2020.

[53] Md Manjurul Ahsan, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md Rahman, Mohammad Shakhawat Hossain, et al. Study of different deep learning approach with explainable ai for screening patients with COVID-19 symptoms: Using ct scan and chest x-ray image dataset. *arXiv preprint arXiv:2007.12525. Unpublished results*, 2020.

[54] Raghavendra Selvan, Erik B Dam, Sofus Rischel, Kaining Sheng, Mads Nielsen, and Akshay Pai. Lung segmentation from chest x-rays using variational data imputation. *arXiv preprint arXiv:2005.10052. Unpublished results*, 2020.

[55] Xi Ouyang, Jiayu Huo, Liming Xia, Fei Shan, Jun Liu, Zhanhao Mo, Fuhua Yan, Zhongxiang Ding, Qi Yang, Bin Song, et al. Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia. *IEEE Transactions on Medical Imaging*, 2020.

[56] Feng Shi, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, Kelei He, Yinghuan Shi, and Dinggang Shen. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE reviews in biomedical engineering*, 2020.

[57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[58] Jared T Hagaman, Ralph J Panos, Gregory W Rouan, and Ralph T Shipley. Admission chest radiograph lacks sensitivity in the diagnosis of community-acquired pneumonia. *The American journal of the medical sciences*, 337(4):236–240, 2009.

[59] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225. Unpublished results*, 2017.

[60] Ahmed Hamimi. Mers-cov: Middle east respiratory syndrome corona virus: Can radiology be of help? initial single center experience. *The Egyptian Journal of Radiology and Nuclear Medicine*, 47(1):95–106, 2016.

[61] Amr M Ajlan, Brendan Quiney, Savvas Nicolaou, and Nestor L Muller. Swine-origin influenza a (h1n1) viral infection: radiographic and ct findings. *American Journal of Roentgenology*, 193(6):1494–1499, 2009.

[62] Kenneth L Bontrager and John Lampignano. *Textbook of radiographic positioning and related Anatomy-E-Book*. Elsevier Health Sciences, 2013.

[63] Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[64] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[65] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[66] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B Blaschko. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 92–100. Springer, 2019.

[67] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[68] Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological methods*, 22(2):322, 2017.

[69] Cheng Chen, Qi Dou, Hao Chen, and Pheng-Ann Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International workshop on machine learning in medical imaging*, pages 143–151. Springer, 2018.

[70] Youbao Tang, Yuxing Tang, Jing Xiao, and Ronald M Summers. Xlsor: A robust and accurate lung segmentor on chest x-rays using criss-cross attention and customized radiorealistic abnormalities generation. *arXiv preprint arXiv:1904.09229. Unpublished results*, 2019.

[71] Jyoti Islam and Yanqing Zhang. Towards robust lung segmentation in chest radiographs with deep learning. *arXiv preprint arXiv:1811.12638. Unpublished results*, 2018.

[72] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410, 2019.

[73] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.