

---

# CT-xCOV: A CT-SCAN BASED EXPLAINABLE FRAMEWORK FOR COVID-19 DIAGNOSIS

---

Ismail ELBOUKNIFY<sup>1\*</sup>, Afaf BOUHOUTE<sup>2</sup>, Khalid FARDOUSSE<sup>2</sup>, Ismail BERRADA<sup>3</sup>, Abdelmajid BADRI<sup>1</sup>

<sup>1</sup> EEA&TI Laboratory, Hassan 2 University, Casablanca, Morocco

<sup>2</sup> Sidi Mohamed Ben Abdellah University, Faculty of Science, Fez, Morocco

<sup>3</sup> Mohammed VI Polytechnic University, School of Computer Science, Benguerir, Morocco

November 27, 2023

## ABSTRACT

In this work, CT-xCOV, an explainable framework for COVID-19 diagnosis using Deep Learning (DL) on CT-scans is developed. CT-xCOV adopts an end-to-end approach from lung segmentation to COVID-19 detection and explanations of the detection model's prediction. For lung segmentation, we used the well-known U-Net model. For COVID-19 detection, we compared three different CNN architectures: a standard CNN, ResNet50, and DenseNet121. After the detection, visual and textual explanations are provided. For visual explanations, we applied three different XAI techniques, namely, Grad-Cam, Integrated Gradient (IG), and LIME. Textual explanations are added by computing the percentage of infection by lungs. To assess the performance of the used XAI techniques, we propose a ground-truth-based evaluation method, measuring the similarity between the visualization outputs and the ground-truth infections. The performed experiments show that the applied DL models achieved good results. The U-Net segmentation model achieved a high Dice coefficient (98%). The performance of our proposed classification model (standard CNN) was validated using 5-fold cross-validation (acc of 98.40% and f1-score 98.23%). Lastly, the results of the comparison of XAI techniques show that Grad-Cam gives the best explanations compared to LIME and IG, by achieving a Dice coefficient of 55%, on COVID-19 positive scans, compared to 29% and 24% obtained by IG and LIME respectively. The code and the dataset used in this paper are available in the GitHub repository [1].

## 1 Introduction

The COVID-19 outbreak has put the medical researchers under an unprecedented spotlight. With the rapid transmission of the virus, identifying effective tools for early diagnosis becomes crucial to prevent further spread. So far, The Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been widely used for detecting COVID-19. However, the difficulties posed by high false negative rates, slow processing times, inconsistent testing protocols, and detection sensitivity in RT-PCR tests [2], present significant obstacles to achieving early and accurate COVID-19 detection. Besides, Chest CT-scans have emerged as a crucial tool for both screening and diagnosing lung infections associated with COVID-19 [3]. Chest CT findings reveal that CT-scans of COVID-19-infected patients show some visual signs such as ground glass opacity, consolidation, and pleural effusion [4]. Therefore, using a CT-scan can enable early detection of COVID-19 infection in patients with initially a negative RT-PCR examination or who are asymptomatic, both before and following the onset of symptoms[2].

With the recent breakthroughs in Artificial Intelligence (AI), more attention has been paid to using AI approaches to assist in making more accurate and reproducible radiology assessments [5]. Several studies have shown the efficacy of AI for accelerating the diagnosis of many diseases. Recently, Deep Learning (DL) models have become increasingly popular for CT-scan imaging-based diagnosis. These models have achieved a remarkable performance, improving

---

\*ielbouknify@gmail.com

the radiologist’s performance in detecting many diseases [6]. However, this increased prediction accuracy is often associated with higher model complexity. The black-box nature of the DL models makes it difficult to comprehend their inner workings, which means that clinicians using DL-based diagnosis will be unable to understand the most important features that contribute to the diagnosis. This issue increases skepticism around the clinical use of DL models and gives rise to a need for interpretability.

eXplainable Artificial Intelligence (XAI) is a new emerging domain, which brings together several processes and methods to provide human users with the ability to comprehend and trust the results of ML algorithms. Recent research in medical imaging analysis is increasingly focused on creating human-understandable explanations of how a model makes specific individual predictions [7]. These explanations are provided as visualization of the image regions that contribute the most to model predictions. While XAI methods have achieved a certain level of success, evaluating and quantifying their effectiveness is still a challenge.

The goal of this paper is to make use of the popular success of DL, the availability of CT images, and the increasing need for interpretability to build accurate and explainable COVID-19 detection systems. To this end, we propose CT-xCOV, an end-to-end framework for explainable COVID-19 diagnosis. The framework adopts popular CNN architectures for lung segmentation and classification, and visual XAI methods for explanations. The CT-xCOV framework is validated using three popular COVID-19 datasets. The key contributions of this paper can be outlined as follows:

- **CT-scan-based COVID-19 detection using Deep Learning (DL):** we use and evaluate different DL architectures for lung segmentation and CT-scan classification. For segmentation, we trained the well-known U-Net model to generate lung masks. For classification, we evaluated different CNN architectures, including DenseNet and ResNet to classify the lungs and distinguish healthy and COVID-19 infected.
- **XAI methods for an explainable detection:** To provide visual explanations about the COVID-19 predictions, we applied three different methods, namely Grad-CAM, Integrated Gradient, and LIME. A textual explanation about the degree of infection, based on this visualization, is also provided.
- **Explainability assessment:** to evaluate the explanations provided by the different methods, we propose a ground-truth-based evaluation method. This method compares the visualization output with the ground-truth infections. The performance is reported as an average of the similarity coefficient between visualization and ground-truth infection masks.

The remainder of this paper is structured into four sections. In section 2, we present an overview of the methodology of our approach and its different steps. In section 4, the performed experiments are detailed, and the obtained results are discussed. Besides we report some related works on COVID-19 detection based on medical images in section 5. Finally, section 6 draws conclusions and gives the perspectives.

## 2 CT-xCOV : An end-to-end framework for eXplainable COVID-19 diagnosis

Following the recent advances in Trustworthy and Explainable Artificial Intelligence (XAI), in this section, we introduce CT-xCOV, our new framework for explainable and automated COVID-19 diagnosis. The section starts with an overview of the CT-xCOV architecture and follows with the research methodology adopted to address our goals.

### 2.1 Architecture Overview

CT-xCOV consists of four key components: 1) data preparation, 2) lung segmentation, 3) COVID-19 detection, and 4) the explainability component. The first component is responsible for collecting, preprocessing, and preparing the CT-scan dataset for the training process. As will be detailed later, to deal with the lack of a large COVID dataset, three different datasets were combined. The second component identifies the Region Of Interest (ROI), i.e. the lung in our case. The third component is in charge of classifying the lungs into two classes, Infected and Uninfected. We performed classification using different CNN architectures. Finally, the last component explains the model prediction and makes it more transparent. To this end, XAI methods are used to provide visual explanations about the model decision. Several XAI visual methods were compared and their performance was assessed. In the case of COVID-19 positive, textual explanation is also provided in forms of infection degree.

Figure 1 illustrates the general architecture and flow of the proposed framework. CT-xCOV is a CT-scan-based framework. A CT-scan volume contains a variable number of 2D axial slices (i.e. the number of slices in a CT scan may vary). Moreover, the shape, as well as the size of the lung, vary significantly between slices. Therefore, CT-xCOV adopts a 2D slice-based approach for COVID-19 detection. The slice-based approach takes as input individual CT slices and generates, for each slice, two outcomes: the infection prediction, and the explanations of the reasons behind this prediction.

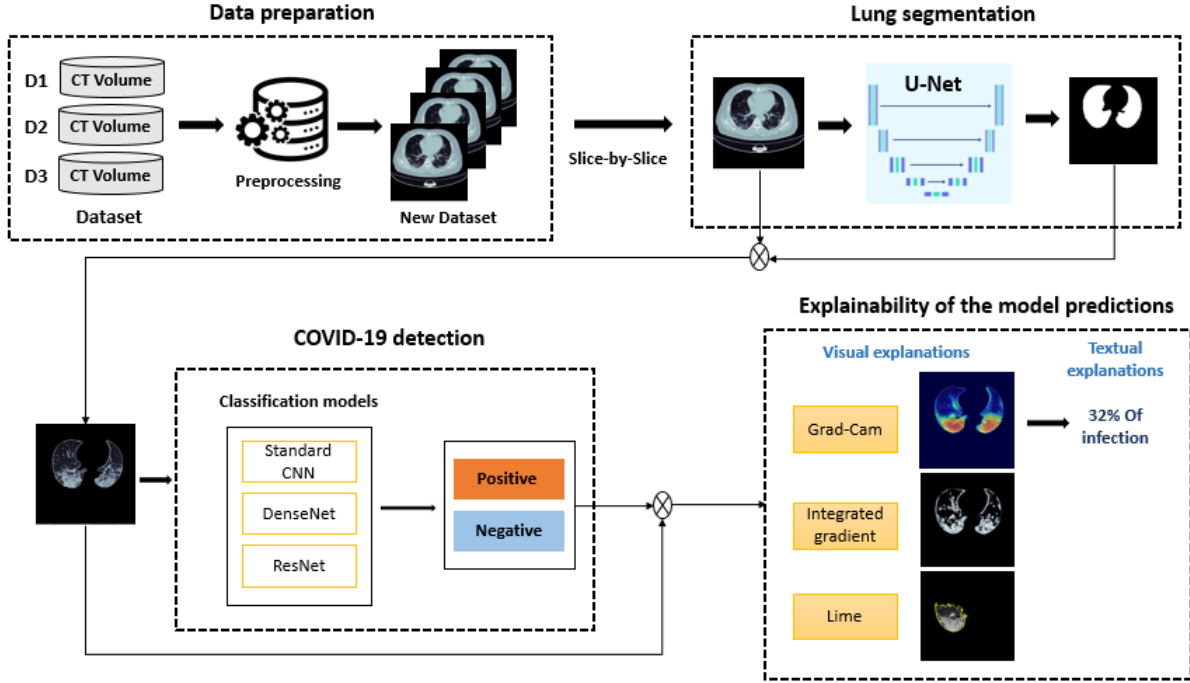


Figure 1: Overall Architecture of the proposed framework.

## 2.2 Data Preparation

This first component of CT-xCOV deals with data preparation. Collected data is pre-processed and prepared considering different pre-processing techniques. As part of our framework design, we considered the following three public datasets:

- **COVID-19 CT Lung and Infection Segmentation Dataset** [8] (referred to as  $D_1$ ). It contains CT volumes of 20 patients, all infected with COVID-19. All the cases were labeled by two radiologists using three labels (left lung, right lung, and infections), and confirmed by a proficient radiologist.
- **COVID-19 CT segmentation dataset** [9] (referred to as  $D_2$ ). It is a dataset of 100 axial CT images from over 40 COVID-19-positive patients. The lungs and infections were segmented by a radiologist.
- **MosMedData** [10] (referred to as  $D_3$ ). It contains CT-scans provided by hospitals in Moscow, Russia. It includes 1110 volumetric CT-scans of unaffected and affected cases. Annotations of infections for 50 case studies were performed by experts from the Moscow Health Care Department.

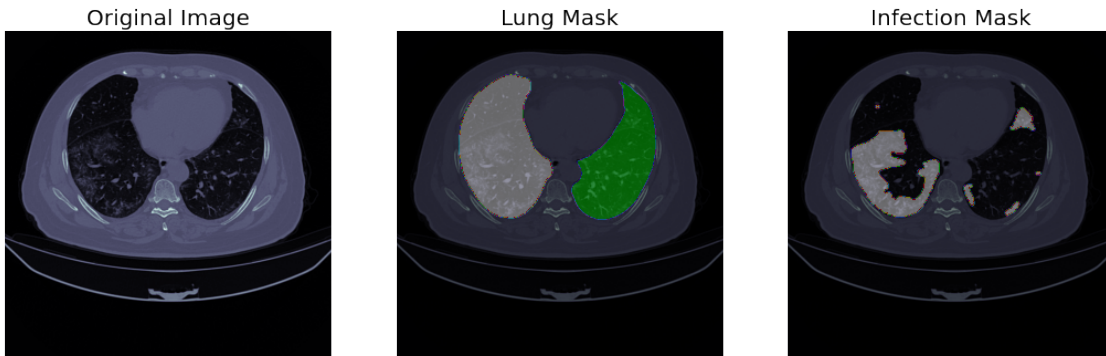


Figure 2: Example of CT-scan from the dataset  $D_1$ , with its corresponding lung and infection masks.

Although the CT-scans in all three datasets are provided in NIFTI (Neuroimaging Informatics Technology Initiative) format, two datasets provide volumetric scans. In contrast, the last one provides axial CT slices (individual slices

belonging to different patients). Therefore, single slices were extracted from volumetric CT-scans. For each volumetric CT-scan with COVID-19, the 20 middle slices were considered. The ground-truth infection masks were used to keep only infected slices. For normal CT-scans, the 5 middle slices were extracted to keep a balanced dataset as the number of normal CTs is higher than that of COVID-19 cases. Finally, all slices were converted to PNG (Portable Network Graphics) images, resized to 256\*256 and their values normalized between 0 and 1.

Dataset	# of Patients Used	# of Images Used	Lung Mask	Lesion Mask
$D_1$	20 Positive	400	✓	✓
$D_2$	>40 Positive	100	✓	✓
$D_3$	50 positive, 254 negative	1770		✓

Table 1: Summary of the datasets used.

Table 1 shows information about the used datasets. To increase the data size and avoid the overfitting problem, helping thus in improving the classification performance, data augmentation was performed. The dataset was augmented by applying different transformation techniques such as horizontal flipping and rotation.

### 2.3 Lung Segmentation

To focus on the lung area, the lung segmentation step is considered. At this stage, U-Net architecture [11], a well-known CNN designed for biomedical applications, is trained to generate lung masks for inputs (CT-scan slices). U-Net has achieved superior performance in many medical image segmentation applications, even with limited datasets. The primary aim of this segmentation step is to extract the lung regions (ROI) for the CTs in the  $D_3$  dataset as it only provides infection masks. For this end,  $D_1$  and  $D_2$  datasets are used to train the segmentation model. This latter takes individual CT slices as inputs and outputs the predicted lung masks. The segmented lung is later fed into the classification model for COVID-19 detection.

### 2.4 COVID-19 Detection

Convolutional Neuron Networks (CNNs) have been commonly used in many aspects of medical image analysis, and have contributed to the great progress in the diagnosis of rapidly evolving diseases. In this prediction component, we compared three different CNN architectures, namely standard CNN (custom) models, DenseNet, and ResNet, for COVID-19 detection. These architectures are briefly presented below.

- ResNet architecture [12] has contributed enormously to the use of very deep neural networks. It limits the gradient loss in the deepest layers by adding a residual bond between each convolution layer. Thus, their main advantages are their ability to reduce gradient leakage in the lower layers of the network, and the possibility of their deep scaling. The latter allows the adjustment of the architecture to the classification task. Although expensive in computational resources, they obtain very good performance in classification on all levels of image detail.
- DenseNet architecture [13] are in line with deep CNNs. Each of the network layers has as input the outputs of the previous layers. This feature increases the number of network connections from  $L$ , for a traditional  $L$ -layer CNN with one connection between each layer, to  $[L(L - 1)]/2$ . Although the DenseNet architecture adopts a similar approach to the ResNet architecture by adding short connections between the layers, it differs significantly in how information is taken from the previous layers. The concatenation of the information made in the DenseNet architecture makes it possible to reduce the number of parameters by a factor of 3 compared to the ResNet architecture while maintaining the same level of performance in terms of [13] precision.
- Our introduced classification model is a standard CNN. It consists of 4 convolution layers, each layer is followed by a max-pooling layer. The role of these layers is the extraction of image features. After that, two dense layers and a Softmax function are used for classification based on the feature extracted from previous layers. Figure 3 shows the CNN architecture considered for classification.

### 2.5 Explainability of the model predictions

To make our framework more transparent and increase the understandability of its decisions, visualization as well as textual explanations are provided. The visual explanations contribute to understanding the model predictions by highlighting the regions in the input CT-scan that were behind the model’s decision, helping to increase confidence

in the model. As part of CT-xCOV, we use three different visual explainability techniques, namely LIME [14], Grad-Cam [15] and Integrated Gradients [16]. The goal is to make available to users different types (gradient-based and perturbation-based) of pixel attribution methods. The three used techniques are described below.

- *Local Interpretable Model-Agnostic Explanations (LIME)* is an XAI technique introduced in [14] to explain how an ML model input features affect its predictions. In the context of image classification, LIME identifies the region of an image, typically a set of super-pixels, that is most strongly associated with a predicted label. To generate explanations, the different steps of LIME are summarized in five steps; (1): Generate random perturbations for the input image. (2): Predict the class for each perturbation. (3): Compute importance weights for each perturbation. (4): Fit an explainable linear model using perturbations, predictions, and weights. (5): Explain the prediction by interpreting the interpretable model.
- *Gradient weighted Class Activation Map (Grad-CAM)* [15] is a visualization technique that is useful for understanding how a CNN was driven to make a classification decision. Grad-CAM uses target gradients in the last convolutional layer to provide explanations by highlighting important regions of an image that impact prediction. The output of Grad-CAM is heatmaps representing the activation classes on the images received as input. An activation map is associated with a specific output class. The class activation map (CAM) is generated by calculating a linear combination of activations, weighted by the weights of the corresponding output for the observed class, This assigns importance to each position  $(x, y)$  in the image based on their contribution to the final prediction.
- *Integrated gradients (IG)* is an explainable AI technique first introduced in [16]. The IG technique aims to provide explanations of how a model’s predictions are influenced by its input characteristics. Its use cases include understanding feature importance, identifying data asymmetry, and debugging model performance. IG has become a popular interpretability technique due to its wide applicability to any differentiable model (text, images, structured data), and its ease of implementation.

In addition to visualization, CT-xCOV also provides textual explanations to report information about the degree of lung infection. It is computed based on the explanation provided by the visualization methods. The degree of infection is calculated as the size of the infected region, highlighted by the visualization, divided by the size of the lungs.

### 3 Experiments

To assess the effectiveness of our CT-xCOV framework, we performed several experiments with different settings. This section describes the experimental setup and defines the metrics used for evaluation.

#### 3.1 Experimental setup

**Segmentation task.** For the lung segmentation component in the proposed framework, we adopted the basic architecture of U-Net [11], with the difference in the input shape becoming  $(256*256)$  and the output shape becoming  $(256*256)$ . U-NET was trained using 554 CT-scan images of 60 patients, and 200 images of 10 patients were kept for testing. Since no lung masks are provided in  $D_3$ , the images used for the segmentation task were mainly taken from  $D_1$  and  $D_2$  datasets. To increase the number of training images, some slices were segmented using the pre-trained model provided by the MedSeg [17] tool. To ensure that the test images are not seen during the training process of U-Net, we performed a train/test split at the patient level. Thus, images from 10 patients were kept for testing. Finally, the training hyper-parameters are summarized in Table 2.

**Classification task.** We evaluated this task using three models. On the one hand, we used DenseNet121 [18] and ResNet50 [19] architectures to extract the image features. Two dense layers and a Softmax function are then employed for classification based on the feature extracted by each architecture. On the other hand, we designed the simple CNN architecture of Figure 3, constituted of 4 layer blocks for feature extraction and two dense layers for classification. Each layer block is composed of 1) a convolution layer with 32 filters of  $3 \times 3$  and ReLu activation function followed by 2) a max pooling layer. For model regularization, a dropout layer is added to the first and last blocks. To ensure that the test images were not seen during the training process, we performed a train/test split at the patient level, where 15% of the images belonging to 50 patients were kept for the test. The training hyper-parameters are summarized in Table 2.

Finally, classification and segmentation models were implemented using Keras [20] and Tensorflow libraries with Python 3.7 on Colab (NVIDIA Tesla K80 GPU, 12 GB RAM) and Kaggle (NVIDIA Tesla P100 GPU, 16 GB RAM).

**Explanation assessment.** The goal of this experiment is to compare and assess the explanations provided by the visualization methods. To this end, we adopt a ground-truth-based evaluation, where the visualization output is

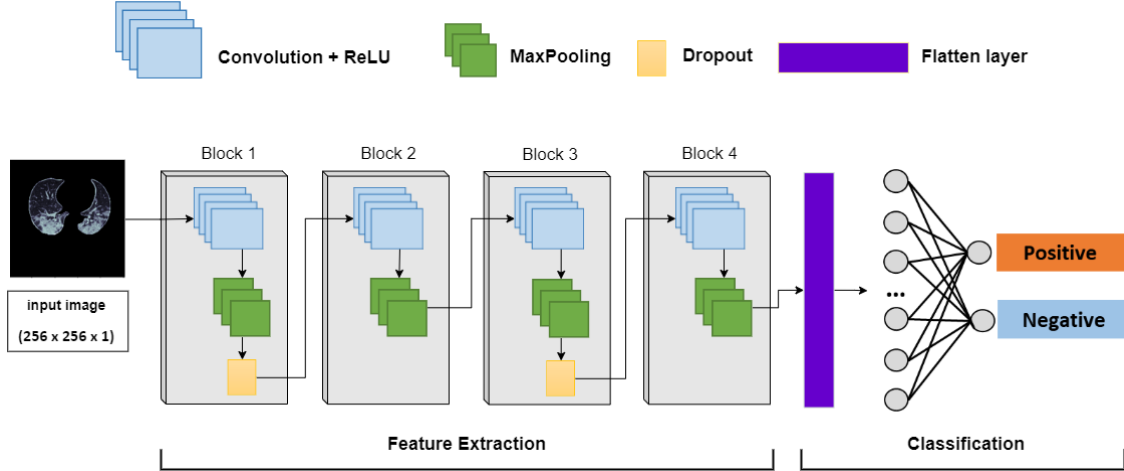


Figure 3: The architecture of our classification model

compared to the ground-truth infections. The comparison is performed by measuring the similarity between the two images. For each CT-scan, we measured the similarity between the output explanations and the ground-truth infection masks. In fact, for the LIME method, we consider the output part of the explanation as the infected region while for heatmap methods (Grad-cam and Integrated Gradient), we applied threshold-based segmentation to extract the highlighted infections. Image thresholding is a well-known image binarization technique, which consists of Converting a grayscale image into a binary image, where the pixel values are restricted to either 1 or 0. Since this thresholding technique depends mainly on the threshold value, different threshold values were tested using the train set to find the value that gives the best segmentation results compared to the infection masks.

Hyper-parameters	Segmentation Model	Classification Model
Batch Size	32	32
Epoch	100	20
Learning rate	0.0005	0.001
Optimizer	Adam	Adam
Loss	Binary Cross Entropy and Dice Loss	Categorical Cross Entropy

Table 2: Hyper-parameters values used in segmentation and classification model

### 3.2 Evaluation metrics

The evaluation of CT-xCOV is performed according to the three considered tasks: lung segmentation, classification, and visual explanations.

**Classification metrics.** We considered different evaluation metrics to report the performance of the evaluated models, namely precision, recall, accuracy, false discovery rate, and F1-score. For classification, prediction outcomes belong to one of the four cases: (\*) True Positive (TP) corresponds to positive samples that are correctly classified as positive, (\*) True Negative (TN) indicates samples that are actually negative and are correctly classified as negative, (\*) False Positive (FP) refers to negative samples incorrectly classified, and (\*) False Negative (FN) refers to positive samples incorrectly classified. Based on these outcomes, evaluation metrics are defined as given below:

- **Accuracy:** is a measure of correct predictions over total predictions:

$$Accuracy = \frac{TP + TN}{N}$$

- **Precision:** measures the proportion of TP among the samples classified as positive:

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

- **Recall:** measures the proportion of actual positives that are correctly identified:

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

- $F_1$  **score:** is a weighted average of Precision and Recall:

$$F_1score = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

- **False discovery rate :** is the portion of FP among all positive predictions:

$$FDR = \frac{FP}{FP + TP}$$

**Segmentation and explanations assessment metrics.** Dice coefficient [21] was used for these tasks. It is a widely adopted statistical indicator used to quantify the similarity between two sample images. Dice measures the spatial overlap between two images A and B as given by the formula below, where the cardinalities refer to the number of pixels. Alternatively, Dice can be expressed in terms of TP, FP, and FN, which refers to the number of pixels in A that are classified correctly or incorrectly concerning the B.

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

## 4 Results

This section discusses the results obtained at each step of our CT-xCOV framework, namely segmentation, classification, and explainability.

### 4.1 Lung segmentation

The Dice coefficient curves for the training and validation of U-Net during the segmentation step are illustrated in Figure 4. Comparing the two curves, it can be noticed that, as the training progresses, the test Dice coefficient follows the train Dice coefficient while remaining slightly inferior. This means that the model’s performance increases over time, so the model is learning. The model achieves a value of 98.33% on the test set.

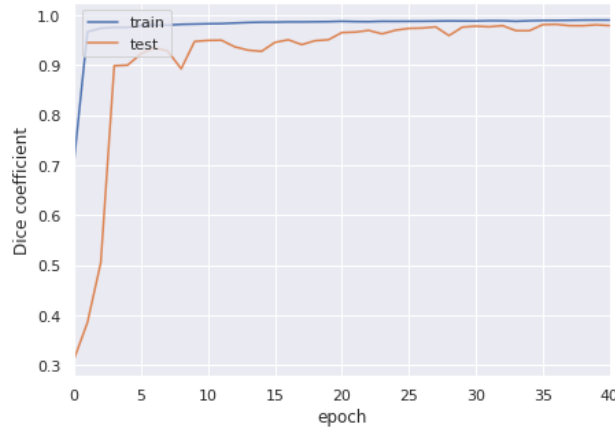


Figure 4: Dice coefficient curves for the U-Net training and test sets.

A visual example of U-Net predictions is shown in Figure 5, which compares an output mask of a random image with the ground-truth mask. We can see that the predicted mask is highly consistent with the ground-truth.

### 4.2 Classification for COVID-19 Detection

The comparison between our proposed CNN model and the other architectures (Resnet and DenseNet) was performed using the 5-fold cross-validation strategy on the train set. The cross-validation results are summarized in Table 3. It can



Figure 5: Example of current mask and mask predicted by U-Net

be observed that the best results are achieved with our proposed CNN. An other difference can be seen in the stability of the models. The performance scores of ResNet50 and DenseNet121 have higher variance (between 3.7% to 12.7%), which means that the predictions of the trained models are unstable. On the other hand, our model achieved higher performance with low variance ( $\leq 1\%$ ).

Model	Precision	Recall	F1-score	Accuracy
Our Model	<b>98.93±0.37%</b>	<b>97.54±0.99%</b>	<b>98.23±0.46%</b>	<b>98.40±0.41%</b>
ResNet50	82.31±14.08%	92.32±2.63%	86.34± 9.22%	85.30±12.60%
DenseNet121	98.42± 1.92%	69.34± 17.69%	80.03±11.34%	85.44±7.35%

Table 3: Comparison between our model and the other architectures.

The model performance on the test dataset is summarized in Table 4. Our model achieved good performance on the test set, with higher scores of accuracy (93%), precision (93%), recall (94%), F1-score (93%), and a lower score of False discovery rate (11.5%).

Score	Precision	Recall	F1-score	Accuracy	FDR
Our Model	<b>93%</b>	<b>94%</b>	<b>93%</b>	<b>93%</b>	<b>11.5%</b>
ResNet50	77%	77%	76%	76%	29.5%
DenseNet121	84%	83%	82%	82%	32.5%

Table 4: Comparison between our model and other CNN architecture on the test dataset

### 4.3 Explainability of the model predictions

To gain insights into the outcomes of our models, we aim to create visual representations of the results. we generate visual explanations using different explainability techniques, namely Grad-CAM, Integrated Gradient (IG), and LIME.

Figure 6 illustrates an example of explanations produced by the three techniques for COVID-19-positive cases, together with the ground-truth infection mask. Comparing the output of the methods, we notice that:

- Grad-CAM (Figure 6c) correctly highlighted the infected parts of the lungs in the CT-scan, which provides an understanding of the model predictions.
- IG (Figure 6d) correctly localized a wide part of COVID-19 infection in the image.
- LIME (Figure 6e) with the top two features obtained using 1000 perturbations, highlighted the regions corresponding to the infected lung area. However, we can see that LIME localized only a part of the infection.

Figure 7 illustrates an example of explanations produced by the three techniques for COVID-19 negative cases, together with the ground-truth infection mask. Comparing the output of the methods, we notice that:

- Grad-CAM (Figure 7b) shows no sign of infection which explains why the model predicted the CT-scan as negative.
- IG (Figure 7c) highlights the contour of the lungs, which may not be as good as other methods' visualizations.



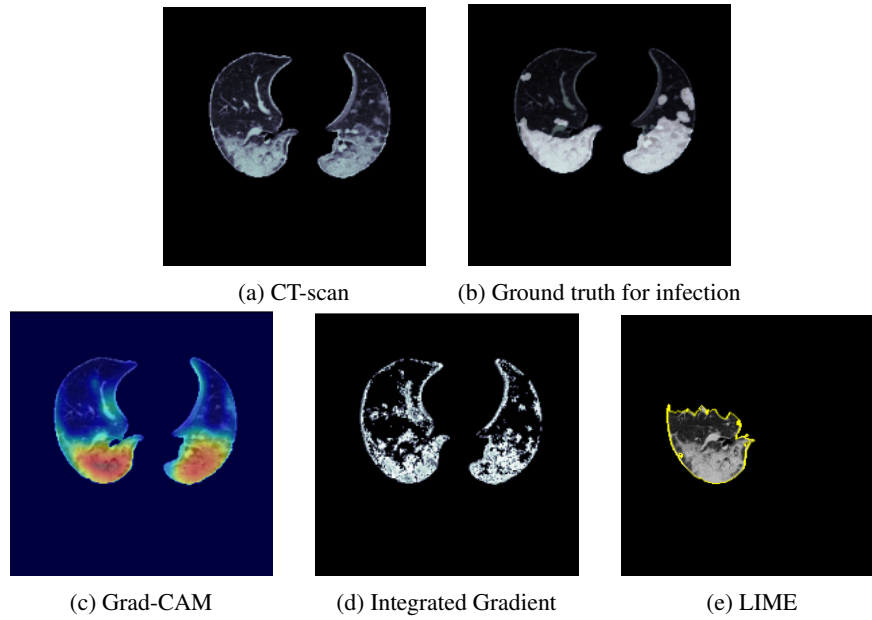


Figure 6: Visualizing positive COVID-19 CT-scan

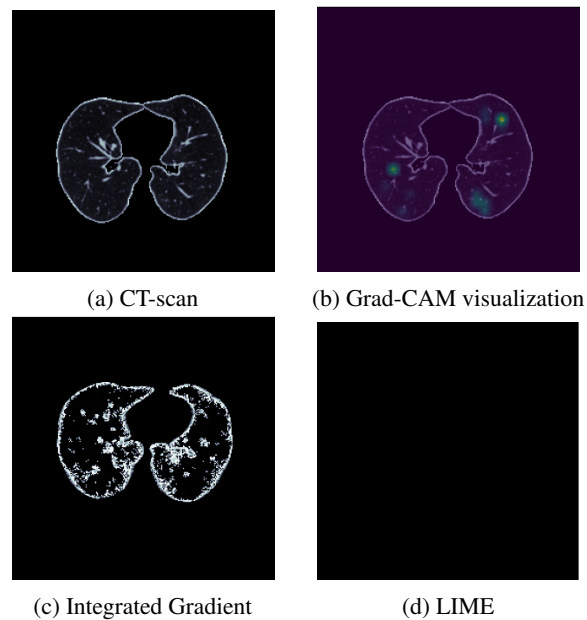


Figure 7: Visualizing negative COVID-19 CT-scan

- LIME (Figure 7d) outputs a black image, which indicates that no infection has been detected.

As mentioned before, CT-xCOV enriches the visualization’s explanations with the textual ones. We provide users with information on the degree of infection, which is calculated as the size of the infected region shown by a visualization method divided by the size of the lungs. Additionally, we provide the degree of infection per lung (i.e. right/left) to give more information about the most infected lung. Figure 8 shows an example of visual and textual explanations using the Grad-Cam method: the infected region is highlighted and the percentage of infection is computed.

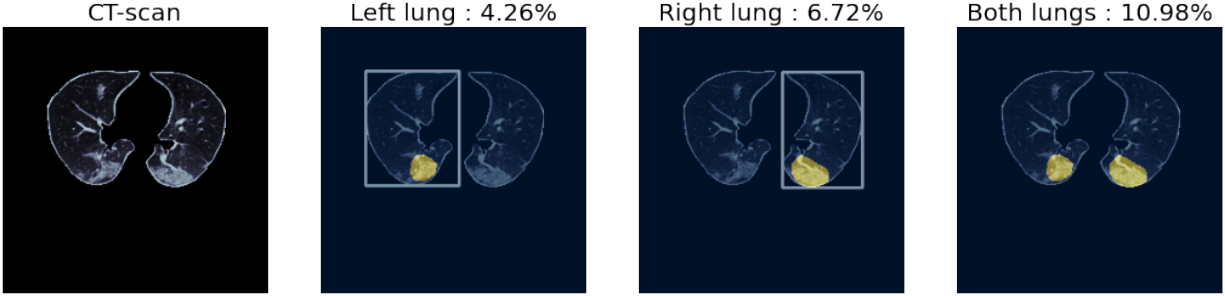


Figure 8: Percentage of infection in the lungs based on the explanation provided by Grad CAM; "Total degree of infection in both lungs: 10.98 %, 4.26 % in the left lung and 6.72 % in the right lung".

#### 4.4 Explanations assessment

As we mentioned before, threshold-based segmentation has been used for segmenting the output visualization. This segmentation depends on the threshold value as illustrated in Figure 9, which shows the infection segmentation obtained by Grad-CAM using different threshold values.

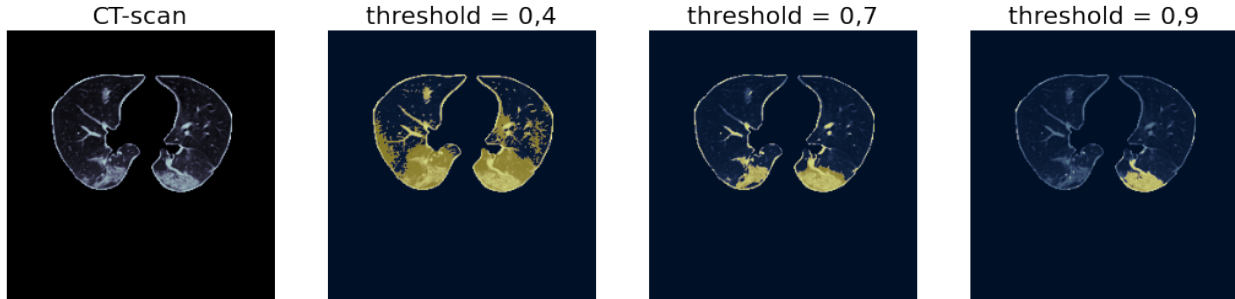


Figure 9: The impact of threshold choice on grad-Cam's explanation

Given the difference between the visualizations provided by the methods in the case of NO-COVID images (IG highlights also the lung contour as a part of the visualization (see Figure 7)), we decided to evaluate the methods for the COVID-19 positive and negative cases separately. Figure 10 shows the dice coefficient plots for the tested threshold values. The Dice coefficients for different threshold values, on COVID-19 positive cases are plotted in Figure 10a. Comparing the plots of the different methods, we can see that the Grad-cam achieved the best result of 0.35 with a threshold of 0.4, followed by the Integrated gradient which achieved 0.19 for a threshold of 0.9. The plots for COVID-19 negative cases are shown in Figure 10b. Comparing the plots, it can be observed that Grad-Cam achieved an average Dice coefficient of 0.81 with a threshold of 0.9, followed by Integrated Gradient for which the average Dice coefficient did not exceed 0.00024 with the threshold of 0.1. Table 5 summarizes the comparison findings for the threshold choice.

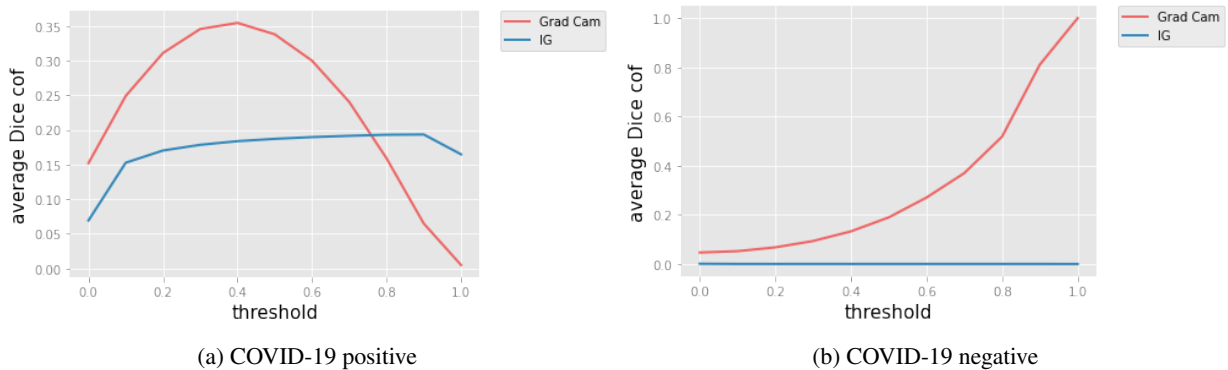


Figure 10: Impact of threshold value on the Dice coefficient for COVID-19 positive and negative cases.

XAI Method	COVID-19 positive cases		COVID-19 negative cases	
	Threshold	Average Dice coeff	Threshold	Average Dice coeff
Grad-cam	0.4	0.35	0.9	0.81
Integrated Gradient	0.9	0.19	0.1	0.00024

Table 5: Best threshold values found on the train set for segmenting Grad-CAM and IG visualizations

After defining the infection segmentation thresholds in the visualization images, the comparison between the XAI methods is performed on the test dataset, using the average of the Dice coefficient as the metric. The results are summarized in the table 6.

XAI Method	Average Dice coefficient (Covid case )	Average Dice coefficient (Normal case )	Explanation Time
Grad-Cam	<b>55%</b>	<b>65%</b>	<b>0.33s</b>
Integrated Gradient	29 %	0.016 %	0.41s
LIME	24 %	77 %	18s

Table 6: Summary of explanations assessment based on the test dataset

The summarized Dice coefficient scores show that:

- For COVID-19 positive scans, Grad-CAM visualizations significantly outperform IG and LIME visualizations by achieving a 55% overlapping with the ground-truth masks versus 29% and 24% achieved by IG and LIME respectively.
- For COVID-19 negative scans, except LIME, higher Dice coefficient values were obtained since we compared visualizations with black images (meaning no lung infection). The low value obtained for IG can be explained by the fact that the IG method produces lung contours leading to high dissimilarity when compared to the black masks.

An additional comparison between the visualization methods was also made by considering the average explanation time (i.e. the time taken to produce the visual explanations). The results obtained are also presented in Table 6. These results show that LIME is time-consuming (an average time of 18s), compared to Grad-CAM and IG which give explanations in less than 1s.

## 5 Related works

This section provides an overview of previous research related to COVID-19 diagnosis. It is divided into two main subsections. In the first subsection, existing approaches for COVID-19 diagnosis based on medical imaging are presented. The second subsection investigates existing works on explainable artificial intelligence methods for COVID-19 diagnosis.

### 5.1 COVID-19 detection based on medical imaging

Several papers propose approaches for detecting COVID-19 infection based on medical imaging. These papers use different techniques, some use DL models ([22]–[24]) and others combine image processing techniques with machine learning models ([25], [26]).

Researchers, recently, are most interested in applying DL techniques for COVID-19 diagnosis. The authors in [22] used a CNN architecture called EfficientNet-b4 to distinguish between infected and normal CT-scan images. The dataset used in this paper consists of 349 CT-scan images of 216 COVID-19 patients and 396 images of non-COVID patients. The dataset was augmented using different transformation techniques such as horizontal/ vertical flipping, and rotation; images were also shifted left, right, upward, and downward. The model performance was evaluated using 5-fold cross-validation. The trained model achieved an Accuracy of 90% and an F1-score of 90%. In another paper [23], the authors aim to accurately detect COVID-19 infection on two types of imaging datasets (CT-scan and chest X-ray). To this end, different DL models such as InceptionResNetV2, VGG16, ResNet50, MobilenetV2, NasNetMobile, DenseNet201, VGG19, and ResNet15V2 have been evaluated. The results demonstrate that NasNetMobile exhibited superior performance in terms of accuracy compared to all other models on both datasets (with scores (81.5% - 95.2%) for CT-scan and (95.4% - 100%) for chest X-ray). In [27], the authors evaluated several DL architectures for COVID-19 diagnosis. Their proposed approach involved utilizing transfer learning with custom-sized inputs optimized for each

specific architecture to attain optimal performance. There are several models included in this paper, such as SqueezeNet, Inception, ResNet, Xception, and DenseNet. They performed a series of experiments on two datasets of CT-scan images, namely, the SARS-CoV-2 CT-scan and COVID-19. The study reported that the evaluated models outperformed previous studies. On the SARS-cov-2 CT dataset, the models achieved a score of 99.4% and 99.8% for accuracy and sensitivity respectively. The scores achieved on the COVID-19 datasets are 92.9% for accuracy and 93.7% for sensitivity.

In [28] the authors involved creating optimized versions of various CNN architectures, including VGG16, VGG19, DenseNet201, Inception\_ResNet\_V2, Inception\_V3, Xception, Resnet50, and MobileNet\_V2, through fine-tuning. To prevent overfitting within their models, they used weight decay and regularizers, known as L2. The performance of these models was evaluated using a dataset comprised of chest X-ray and CT scans, for binary classification. According to the findings, Resnet50, MobileNet\_V2, and Inception\_Resnet\_V2 which were fine-tuned demonstrated exceptional performance with a considerable increase in both training and validation accuracy, achieving more than 96% accuracy. However, CNN, Xception, VGG16, VGG19, Inception\_V3, and DenseNet201 had relatively low performance, with an accuracy of no more than 84%.

Other works combine image processing techniques with machine learning models for COVID-19 detection. Authors of [25] use Histogram of Oriented Gradients (HOG) to extract features in chest X-ray images of COVID-19. They compared the accuracy of several machine learning classification models, namely SVM, KNN, NB algorithm, and DT. The results demonstrated that the SVM is more accurate than other machine learning classification models. In [26], the authors apply two DL models (DenseNet and GoogleNet) for feature extraction and stacking ensemble learning technique to COVID-19 diagnosis, using an X-Ray dataset. The result indicated that the final accuracy of the proposed model was good compared to previous work.

## 5.2 Explainable AI for COVID-19 Diagnosis

Lately, much research on automatic COVID-19 diagnosis has been oriented toward creating human-understandable explanations of how DL models make specific individual predictions. In [23], authors use LIME to explain their classification model. A weighted linear model was employed, which determined a coefficient for each super-pixel in the image. These coefficients indicate the level of impact that each super-pixel has on the prediction of COVID-19 infection. After this, the top features were ranked to identify the most significant super-pixels. In [27], The authors utilized two visualization techniques to offer visual explanations of the predictions made by their models, namely t-distributed stochastic neighboring embedding (t-SNE) and Grad-CAM. The visual representations suggest that the models accurately pinpoint the areas linked to COVID-19. In [29], the authors propose an explainable COVID-19 classifier for CT-scan dataset. The proposed approach includes an XAI module that can offer supplementary diagnostic information to radiologists. To perform their approach, CT-scan data was collected from hospitals in China. The collected data consists of 380 CT-scan with COVID-19 infection and 424 negative CT-scan. To ensure a just evaluation, models were trained on a private dataset and then tested on the open access publicly available CC-CCII dataset, which contains 2034 CT-scan. The proposed XAI module provides visualization based on CAM (Class Activation Map). Moreover, The adoption of the LIME and Shapley methods was aimed at interpreting the individual contributions of each super-pixel and examining the differences between local and global explainability. By utilizing these XAI modules, it was possible to identify the areas within the CT images that are more indicative of lesion presence, thereby enabling the model to arrive at a clear final decision.

The above discussion of related research demonstrates the crucial role of artificial intelligence in addressing the challenge of automated COVID-19 detection, as well as the increasing interest in interpretability/explainability to provide trustworthy detection systems. Table 7 presents a summary of the discussed works. For the sake of simplicity, for the works evaluating several DL models, we reported the results of the best model. Despite the success of ML and DL methods, which have proven to be very powerful for medical image classification, and the introduction of XAI methods, challenges in evaluating the effectiveness of XAI methods and in providing understandable explanations still remain.

## 6 Conclusion

In this paper, we propose CT-xCOV, an end-to-end approach for explainable COVID-19 diagnosis using CT-scans. The framework consists of four main components: (1) Data preparation, (2) lung segmentation, (3) COVID-19 detection, and (4) Explainability.

The proposed CNN manages to achieve accuracy and an F1-score of 98% and is more stable compared to the two evaluated models. The visual explanations, obtained by the different XAI methods tested, show that the proposed CNN model has correctly detected COVID-19-related regions as specified in the infection masks. Such explanations are very

Ref	Dataset	Classification approach	Performance	Interpretation
[22]	COVID-CT-Dataset [30]	EfficientNet b4 [31]	- Accuracy=90% - F1-score=90%	–
[23]	CT and X-ray images collected from Kaggle	NasNetMobile	Accuracy on: - CT (81.5%–95.2%) - X-ray (95.4%–100%)	LIME
[25]	X-ray images [32]	XGBoost	- Acc=97.87% - F1-score=97.87%	–
[27]	SARS-CoV-2-CT-scan [9]	ResNet101	- Acc=99.4% - F1-score=99.4%	Grad-CAM
	COVID19-CT [24]	DenseNet	- Acc=92.9% - F1-score=92.9%	
[29]	CC-CCII dataset [33]	XAI based classifier	Acc=89.23% AUC=93.22%	CAM, LIME and SHAP

Table 7: Comparative table of most relevant related works.

useful to create a relationship of trust between the DL model and the experts in the field. Given the lack of standard metrics for evaluating XAI methods, we considered a ground-truth-based evaluation method using infection masks to compare the visualization performance of these methods. This evaluation consists of comparing and measuring the similarity between generated visualization and ground-truth infection masks using the well-known Dice coefficient metric. This comparison shows that Grad-CAM achieves good results in terms of the Dice coefficient (56% on positive cases, and 65% on negative cases) and explanation time (0.33s). In future work, we intend to implement and investigate other metrics for evaluating interpretability, which will make a huge contribution to the growth of this line of research. We also intend to compare explainability achieved through DL-based approaches with approaches using interpretable ML approaches combined with feature extraction. The aim of this is to create clear and defined features from CT-scan images using traditional image analysis techniques.

## References

- [1] *Ct-xcov*. [Online]. Available: <https://github.com/ismailelbouknify/CT-xCOV>.
- [2] S. Hu, Y. Gao, Z. Niu, *et al.*, “Weakly supervised deep learning for covid-19 infection detection and classification from ct images,” *IEEE Access*, vol. 8, pp. 118 869–118 883, 2020.
- [3] Y. Fang, H. Zhang, J. Xie, *et al.*, “Sensitivity of chest ct for covid-19: Comparison to rt-pcr,” *Radiology*, 2020.
- [4] T. C. Kwee and R. M. Kwee, “Chest ct in covid-19: What the radiologist needs to know,” *Radiographics*, vol. 40, no. 7, p. 1848, 2020.
- [5] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, “Artificial intelligence in radiology,” *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.
- [6] Y. Sun, Y. Qu, D. Wang, *et al.*, “Deep learning model improves radiologists’ performance in detection and classification of breast lesions,” *Chinese Journal of Cancer Research*, vol. 33, no. 6, p. 682, 2021.
- [7] H. Hakkoum, I. Abnane, and A. Idri, “Interpretability in the medical field: A systematic mapping and review study,” *Applied Soft Computing*, vol. 117, p. 108 391, 2022.
- [8] *Zenodo*. [Online]. Available: <https://zenodo.org/record/3757476#.YqBY6KjMLIX>.
- [9] *Medical segmentation*. [Online]. Available: <http://medicalsegmentation.com/covid19/>.
- [10] S. P. Morozov, A. Andreychenko, N. Pavlov, *et al.*, “Mosmeddata: Chest ct scans with covid-19 related findings dataset,” *arXiv preprint arXiv:2005.06465*, 2020.
- [11] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [14] M. T. Ribeiro, S. Singh, and C. Guestrin, ““ why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

- [15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [16] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*, PMLR, 2017, pp. 3319–3328.
- [17] *Medseg*. [Online]. Available: <https://htmlsegmentation.s3.eu-north-1.amazonaws.com/index.html>.
- [18] G. Huang, Z. Liu, and K. Q. Weinberger, “Densely connected convolutional networks,” *CoRR*, vol. abs/1608.06993, 2016. arXiv: [1608.06993](https://arxiv.org/abs/1608.06993). [Online]. Available: <http://arxiv.org/abs/1608.06993>.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *CoRR*, vol. abs/1512.03385, 2015. arXiv: [1512.03385](https://arxiv.org/abs/1512.03385). [Online]. Available: <http://arxiv.org/abs/1512.03385>.
- [20] *Keras*. [Online]. Available: <https://keras.io/api/applications/>.
- [21] I. R. I. Haque and J. Neubert, “Deep learning approaches to biomedical image segmentation,” *Informatics in Medicine Unlocked*, vol. 18, p. 100297, 2020.
- [22] T. Anwar and S. Zakir, “Deep learning based diagnosis of covid-19 using chest ct-scan images,” in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, IEEE, 2020, pp. 1–5.
- [23] M. M. Ahsan, K. D. Gupta, M. M. Islam, S. Sen, M. Rahman, M. S. 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*, 2020.
- [24] X. He, X. Yang, S. Zhang, *et al.*, “Sample-efficient deep learning for covid-19 diagnosis based on ct scans,” *medrxiv*, 2020.
- [25] N. Khan, F. Ullah, M. A. Hassan, A. Hussain, *et al.*, “Covid-19 classification based on chest x-ray images using machine learning techniques,” *Journal of Computer Science and Technology Studies*, vol. 2, no. 2, pp. 01–11, 2020.
- [26] M. Fauvel, J. Chanussot, and J. Benediktsson, “A combined support vector machines classification based on decision fusion,” in *2006 IEEE International Symposium on Geoscience and Remote Sensing*, IEEE, 2006, pp. 2494–2497.
- [27] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, “Explainable covid-19 detection using chest ct scans and deep learning,” *Sensors*, vol. 21, no. 2, p. 455, 2021.
- [28] K. El Asnaoui, Y. Chawki, and A. Idri, “Automated methods for detection and classification pneumonia based on x-ray images using deep learning,” in *Artificial intelligence and blockchain for future cybersecurity applications*, Springer, 2021, pp. 257–284.
- [29] Q. Ye, J. Xia, and G. Yang, “Explainable ai for covid-19 ct classifiers: An initial comparison study,” in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2021, pp. 521–526.
- [30] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, “Covid-ct-dataset: A ct scan dataset about covid-19,” *arXiv preprint arXiv:2003.13865*, 2020.
- [31] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
- [32] J. P. Cohen, P. Morrison, and L. Dao, “Covid-19 image data collection,” *arXiv preprint arXiv:2003.11597*, 2020.
- [33] K. Zhang, X. Liu, J. Shen, *et al.*, “Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography,” *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.