Explainable Deep Learning for Covid-19 detection using chest X-ray and CT images

Sidi Ahmed Mahmoudi and Sédrick Stassin and Mostafa El Habib Daho and Xavier Lessage and Saïd Mahmoudi

Abstract Recently, the applications of Artificial Intelligence (AI) and more particularly Deep Learning (DL) have gained high importance in several domains such as computer vision, robotics, medical imaging, etc. Despite their excellent results in terms of precision, the behaviors and decisions of these AI and DL algorithms are not always explainable and interpretable, which make from them a black box. Since May 2018, the general data protection regulation (GDPR) requires a right of explanation for the output of an algorithm, which is necessary and justified for several examples such as autonomous cars and computer-aided diagnosis (CAD) systems. As result, a high interest in terms of research has been given recently to the domain of Explainable Artificial Intelligence (XAI). In this paper, we propose an approach for explaining Deep Learning algorithms when applied to images classification and segmentation. The proposed approach allows to provide the most appropriate explanation method and the most accurate and explainable DL model. As use case, we applied our approach for explaining DL models used for covid-19 images classification and segmentation with two modalities: X-ray images and CT scans. Experimental results showed the interest of our explanation approach within three facts: 1) identification of the most interpretable DL model; 2) measure of positive and negative contribution of input parameters (image pixels) in the decision of DL models; 3) detection of data (training and validation datasets) biases, where

Sidi Ahmed Mahmoudi
University of Mons, 20, Place du Parc, 7000, Mons, Belgium. e-mail: sidi.mahmoudi@umons.ac.be

Sédrick Stassin
University of Mons, 20, Place du Parc, 7000, Mons, Belgium. e-mail: sedrickstassin@hotmail.com

Mostafa El Habib Daho
University of Tlemcen, Algeria. e-mail: mostafa.elhabibdaho@univ-tlemcen.dz

Xavier Lessage
University of Mons, 20, Place du Parc, 7000, Mons, Belgium. e-mail: xavier.lessage@umons.ac.be

Saïd Mahmoudi
University of Mons, 20, Place du Parc, 7000, Mons, Belgium. e-mail: said.mahmoudi@umons.ac.be

the deep neural networks are focusing on image regions that are not supposed to be important. The provided explanations were evaluated by doctors and physicians that confirmed that accuracy of our results.

1 Introduction

During the last years, advances in Artificial Intelligence (AI) have gained a great importance in our daily lives within several domains such as computer vision, robotics, medical imaging, medicine, etc. If we follow the history of Artificial Intelligence, we can distinguish two forms of AI: symbolic programming and machine learning. The symbolic approach allows to represent the human knowledge in a declarative and sequential form using facts, rules and conditions (such as: if-so rule) allowing to illustrate all situations. The Machine Learning (ML) approach consists of developing models that are able to mime and learn information from data and examples in order to provide a generalized solution for unknown examples. In this context, Deep learning presents an important branch of Machine Learning, which proposes deep neural network architectures that are composed of multiple layers allowing to transform annotated data (input) into a representative model of data. In the domain of computer vision and medical imaging, input data (first layer) are presented by a matrix of pixels. The other layers allow to detect and combine features (corners, edges, faces, etc.) in order to classify images, localize objects or segment images. Notice that convolutional neural networks (CNNs) [1] are particularly used in the domain of computer vision and medical imaging , where images features are calculated within the application of different convolutions. The success of Deep learning is mainly due to the advances of the domains of high-performance computing (HPC) and the high availability of massive volumes of data (Big Data). This allowed to execute deep neural networks that are composed of tens or even hundreds of layers with a huge number of connections between neurons, increasing the number of adjustable parameters to hundreds of millions. In terms of precision and accuracy, these networks solved and even outperformed human is several tasks such as video games, images classification and retrieval, car driving, etc. However, these AI algorithms and more particularly Deep Learning models are considered as black boxes since they and not easily explainable and interpretable. In fact, the proposed solution given by the neural network, leads us to ask questions about the path and rules taken to achieve this result. This is commonly called the black-box problem of neural networks. Even if we create a network using training, validation and test datasets, we have no idea what the network detects exactly, and what makes it ultimately choose. So, do we trust the DL models decisions? In this context, model interpretation allows to understand and explain these decisions by the response function, i.e., the what, why, and how? In this paper, we propose an approach for explaining Deep Learning algorithms when applied to images classification and segmentation such as required in the problem of Covid-19 detection using medical images. The proposed approach allows to provide the most appropriate explanation method (perturbation-based approach, gradient-based approach, relevance-based approach and proxy models) and the most accurate and explainable DL model by following three main steps: 1) analysis and comparison between XAI visualization methods for the predicted class only; 2) comparison and analysis of XAI visualization methods between all the existing classes; 3) Non-visual evaluation of XAI methods accordingly to noise injection. As use case, we applied our explanation approach for covid-19 images classification and segmentation using two modalities: X-ray images and CT scans. Experimental results showed the interest of our explanation approach within three facts: 1) identification of the most interpretable DL model; measure of positive and negative contribution of input parameters (image pixels) in the decision of DL models; 3) detection of data (training and validation datasets) biases, where the deep neural networks are focusing on less important regions.

The remainder of this paper is organized as follows: Section 2 outlines the related works in the domain of explainable Deep Learning and more particularly those applied for images classification and covid-19 detection and segmentation. In Section 3, we describe our Deep Learning approach for covid-19 detection and segmentation using X-ray and CT images. The fourth section is devoted to present our DL explanation approach, while Section 5 presents experimental results. Finally, conclusions and future works are presented in the last Section.

2 Related Work

In literature, we can categorize two kinds of works related to our approach of explaining image classification Deep Learning models that are used for covid-19 detection: Deep Learning explanation approaches and covid-19 deep learning detection models

2.1 Deep Learning explanation approaches

The explanation and interpretation of deep neural networks is necessary to solve the black box problem. It is also necessary to have a complete and correct explanation.

Gilpin et al. [2] presented two possible ways to evaluate an explanation : interpretability and completeness of the explanation, where interpretability is defined by "the ability to explain in comprehensive terms to a human" [3]. On the other hand, the completeness allows to describe the working mechanism of a system in an accurate way. An important challenge in Explainable Artificial Intelligence is present in the required trade-off between interpretability and completeness. The most complete explanation of a deep neural network can always be the description of its mathematical functioning, which is not easily interpretable, and will not be a good explanation for everyone. On the contrary, the easiest explanation will never be complete.

In literature, one can find several methods, published recently, that trend to explain deep neural networks. In this paper, we focus our research on classification deep neural networks in order to provide an adapted explanation for our use case problem: covid-19 detection and classification using X-ray and CT images. The main related works in this area are focusing on explaining convolutional neural networks (CNN) since they are mainly used for computer vision and medical imaging applications. We can categorize four main explanation approaches: perturbation-based methods, gradient-based methods , relevance-based methods and proxy models.

2.1.1 Perturbation-based methods

1. **Feature ablation:** involves the replacement of each feature or a group of features by a baseline value in order to compute the output difference. A low difference means that the replaced features are less important and vice versa.
2. **Feature permutation:** Molnar [4] presented similar approach to the ablation one, where features are switched between them within a batch. A feature is considered as important if the shuffling causes an increase of model error and vice versa.
3. **Occlusion:** mainly used in image classification and developed by Zeiler and Fergus [5], which proposed to replace a square of input pixels by a grey square. The occluded pixels are important if the class probability drops significantly. This method can be seen as an application of feature ablation to image classification.

The main drawback of the perturbation-based approaches is the highly intensive computation since the output needs to be recomputed after each new applied perturbation. The computation gets more intensive as the input image gets larger.

2.1.2 Gradient-based methods

1. **Deconvolutional networks:** Zeiler & Fergus [5] proposed an approach using a deconvolutional network to visualize the most discriminating parts of an image. The deconvolutional layer is created for each convolutional layer, providing a path back. This approach is based on five steps (Fig. 1):

   - an image is feeded to the trained model
   - the model computes the forward pass up to the last convolutional layer (or another chosen convolutional layer)
   - The strongest activation (or a selected activation) of this layer is left non-zero
   - The reverse order of the operations carried out during the forward pass is executed by unpoolying, rectifying, and filtering until the input of the model is reached.
   - a reconstructed image shows what strongly activates the input image with the current model

   Notice that the network starts by extracting basic features such as edges and corners in the first convolutional layers. Then, it extracts general shapes and ends with the recognition of objects to be classified.

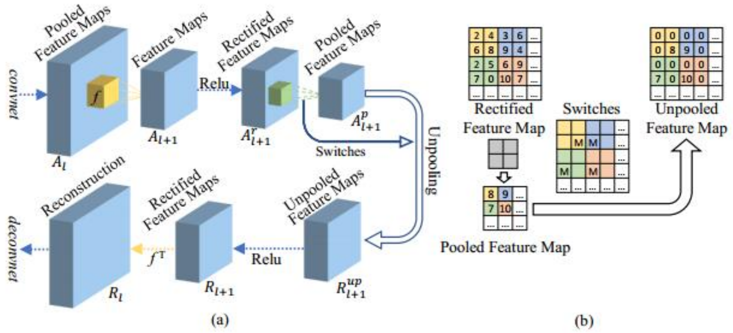


Fig. 1: (a) Top : classical operations in a convolutional network (filtering, rectifying, pooling); Bottom : associated deconvolutional operations (filtering, rectifying, unpooling); (b) Illustration of the unpooling operation [6]

2. **Gradient (or backpropagation):** based on the deconvolutional method from Zeiler & Fergus [5], Simonyan et al. [7] proposed a new approach that can be described as follows :

   If we consider I_0 as a given image, c as a class and $S_c(I)$ as a class score function for a classification problem to be solved with a convolutional network. The goal is to rank the pixels of I_0 upon their impact on $S_c(I)$. By using the derivative $\frac{\partial S_c}{\partial I}|_{I_0}$, the pixels importance can be computed. This backpropagation is applicable to any layer (dense layers for example) whereas deconvolutional network is only applicable to convolutional layers.
3. **Guided Backpropagation:** proposed by Springenberg et al. [8], which combine the convolution and backpropagation method with *rectifying* : if at least one of the entries values compared to the top gradient **or** bottom signal data are negatives, it will be masked (row 4 of Fig. 2 compared to row 2 and 3). This is the only significant difference from previous methods. It is called *guided* because it has another guidance from the higher layers compared to classic backpropagation. It erases the backward flow of negative gradients, thus reducing the activation of the higher layer unit we want to visualize.
4. **Class Activation Mapping (CAM)** proposed by Zhou et al. [9], which highlights the most discriminative image regions for a chosen class. Based on the fact that convolutional layers retain spatial information, and that higher-level visualizations are represented by the last convolutional layers of CNNs, the best choice for visualization is the last convolutional layer. The neural network must be composed of a sequence of convolutional layers, followed finally by a GAP (Global Average Pooling), which will use the features extracted from the last convolutional layer for a fully connected layer, giving the probabilities by class.

   For a selected image, let $f_k(x, y)$ be the activation of the unit k, at location (x, y) in the last convolutional layer of the network. The global average pooling

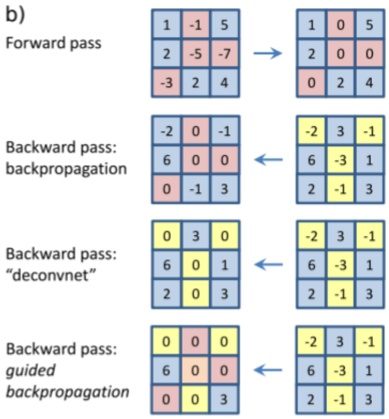


Fig. 2: Different methods of propagating back through a ReLU nonlinearity [8]

operation is represented by $F^k = \sum_{x,y} f_k(x, y)$. For a class c, the probability is calculated as $S_c = \sum_k w_k^c F_k$ where w_k^c is the weight corresponding to class c for unit k. By replacing F_k into S_c, we obtain

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y) \tag{1}$$

Finally, M_c is the class activation map for class c, where each spatial element is given by

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{2}$$

As a result, $S_c = \sum_{x,y} M_c(x, y)$, and $M_c(x, y)$ indicates the importance of the activation at spatial grid (x, y) leading to classify an image to class c.

5. **Gradient-Weighted Class Activation Mapping (Grad-CAM)**: proposed by Selvaraju et al.[10] representing an improvement of CAM method [9], Where CAM is only applicable to CNN without fully connected layers. Grad-CAM can be used with fully connected layers, and therefore for a broader range of CNNs. Grad-CAM uses the gradient information flowing into this layer to compute the importance of each feature map from this last convolutional layer for the predicted class.

Other explanation methods exist in literature, which are also based on the gradient such as "Gradient*input" [11], "Integrated gradient" [12], "SmoothGrad" [13] and Guided Grad-CAM by Selvaraju et al. [10].

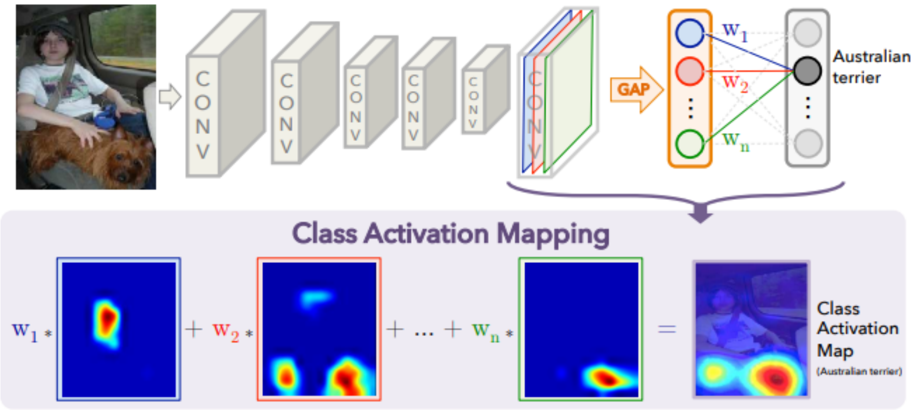


Fig. 3: CAM Method [9]

2.1.3 Relevance-based methods

1. **Layer-Wise Relevance Propagation (LRP)**: the layer-wise relevance propagation method is a conservative back-propagation technique that uses several purposely designed rules, created by Bach et al. [14]. The conservative property is ensured in this way : what has been received by a neuron must be redistributed to the lower layer in equal amount, this is true for any layer (Fig. 4). A neuron's weight from the final layer (a class probability for instance) back-propagated to the input layer will have his weight summed by the neurons of any layer. Several approaches based on LRP are proposed such as : LRP-Z (or LRP-0), LRP-ϵ, LRP-$\alpha\beta$, LRP-Flat and LRP-Preset.

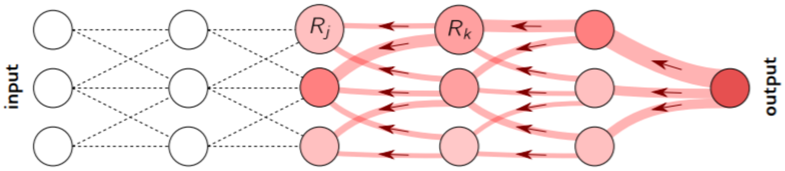


Fig. 4: Illustration of the LRP procedure. Each neuron redistributes to the lower layer as much as it has received from the higher layer [15]

2. **Deep Taylor Decomposition**: it exploits the structure of the neural network with the propagation of explanation from the output to the input layers using a predefined set of rules [16].

2.1.4 Proxy models

Proxy models allow to reduce complexity of deep neural networks (such as ResNet, NasNetLarge, etc.) and other classifiers. A proxy model has a similar behavior to the initial model, but is easier to explain.

1. **Linear model : LIME**: proposed by Ribero et al. [17], which stands for *Local Interpretable Model-agnostic Explanations*. Local means that the provided explanation will be specific to an input image feeded to the model, and will not probe the global model behavior. Interpretable specifies that a qualitative understanding between the input and the prediction is needed. This interpretation depends on the audience, on the user limitations, and therefore should be easy to understand. Finally, model-agnostic relates to the applicability to any classifier that exists or even will be created in the future. Therefore, this technique does not rely on the inner principles of neural networks (layers, units).
2. **Decision Tree**: decision trees allow to classify an instance, beginning at the root of the tree. On each node, a test is applied to design the the convenient branch where a class is assigned by the leaf reached. In the 1990s, research work has been made to decompose shallow neural networks into decision trees, as they are much more interpretable. Now, with the arrival of deep neural networks, new techniques needed to emerge to generalize to the hidden layers.
3. **DeepRED**: proposed by [18], which uses a decision tree as a proxy model.

2.2 Covid-19 Deep Learning models

Despite the short period of time since the appearance of COVID-19, several studies have been carried out to detect this disease from X-ray and CT images of the chest. Different deep learning-based architectures are used for accurate disease detection [19] [20] [21] [22]. Other researchers have been interested in proposing explainable architectures to convince doctors of the decision of their models. Among them, we find the work of [23] where the authors proposed a transfer learning approach using the CheXnet model [24]. The obtained Densenet-121 model was trained and tested on a public dataset containing 13,800 chest radiography images across 13,725 patients. The authors have also performed an interpretability analysis using Grad-CAM [25] to highlight the most important image regions in making a prediction. Their model achieved an average accuracy of 92.91% using patient-wise k-fold cross-validation. In [26], authors fine-tuned the SqueezeNet architecture with Bayesian optimization and data augmentation. They generated visual explanations of their model decisions using class activation mapping. The proposed approach was trained and tested on 5949 posteroanterior chest radiography images for 2839 patient cases, the accuracy of the proposed model reached to 98.3%. In [27], authors proposed a deep convolutional neural network-based architecture, named as CovXNet, that used depth wise convolution with varying dilation rates for efficiently extracting diversified features from chest X-rays. In their experiments, different forms of CovXNets are designed and trained with X-ray images of various resolutions and for further optimization of their predictions, a stacking algorithm is employed. Finally, a gradient-based discriminative localization using Grad-CAM is integrated to distinguish the abnormal regions of X-ray images referring to different types of pneumonia. They obtained an accuracy rate of 90.2% for multiclass classification (COVID, normal, Viral and Bacterial pneumonias). In [28], authors proposed an explainable deep neural network, called DeepCOVID Explainer where the classification is made using a combination of three models: VGG-19, ResNet-18, and DenseNet-161. To improve the COVID-19 detection transparency, class-discriminating regions on the subject's chest are generated by employing Grad-CAM [25], Grad-CAM++ [29], and LRP [30]. They obtained a positive predictive value (PPV) of 89.61% and recall of 83% using 16,995 Chest X-Ray images across 13,808 patients, covering normal, pneumonia, and COVID-19 cases. In [31], authors proposed and evaluated an approach based on transfer learning exploiting the VGG-16 model. They built two models, the first one aimed to detect whether a chest X-ray is related to a healthy patient of to a patient with generic pulmonary disease and the second one distinguishes between the COVID-19 infection and the other pulmonary diseases. Moreover, to provide explainability, they proposed to visualize class activation maps using the Grad-CAM algorithm. The experimental results considered two different datasets for a total of 6,523 chest X-rays, showing an accuracy of 96% for the first model, and an accuracy of 98% for the COVID-19 detection.

3 Covid-19 detection and classification

This section is presented within two main parts: covid-19 detection using x-ray images and covid-19 detection using CT images.

3.1 Covid-19 detection using X-ray images

Before staring the process of models explanation, we start by their development with four steps: X-ray data collection, data augmentation, transfer learning and models evaluation.

3.1.1 X-ray data collection

The dataset used in this study was proposed by [32], where images are collected from three different sources:

- The Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 DATABASE [33]

- Novel Corona Virus 2019 Dataset developed by Joseph Paul Cohen and Paul Morrison and Lan Dao in GitHub [34]
- Images extracted from 43 different publications

References of each image are provided in the metadata. Normal and Viral pneumonia images were adopted from the Chest X-Ray Images (pneumonia) database [35]. All the images are in Portable Network Graphics (PNG) file format and resolution is 1024-by-1024 pixels.

3.1.2 Data augmentation

in order to increase our dataset and get a better accuracy, we used a common practice in deep learning, called "data augmentation". This practice has the effect of presenting information in different aspects of the image, which is favorable to the training of the parameters and avoids overfitting. Our augmentation strategy is purely geometric and is applied during training. For each input image, we applied a random combination of operations to provide the network with variations in the information present in the original images at each iteration. The operations used are rotation, zoom, horizontal flip, and rotation. The parameters are random and chosen according to a fixed interval. This process of data augmentation allowed to increase the size of our dataset by a ratio of 30%.

3.1.3 Transfer Learning

In order to provide accurate results, we propose to exploit the technic of transfer learning using pre-trained classification models, where the weights are initialized with the ImageNet database [1]. This allows to benefit from previously acquired learning weights for solving another classification problem (covid-19 images classification). The pre-trained CNN models are composed of multiple layers that allow to transform input annotated data into a representative model of data. For images classification, the input layer is represented by a matrix of pixels and the other layers allow to analyze and combine pixels values in order to detect specific features such as corners, edges, faces, etc. The last layer is dedicated to predict the corresponding class to the input image. During the learning phase, the initialized weights are updated after each epoch (iteration) in order to classify covid-19 images (3 classes) instead of ImageNet images (1000 classes). This process allows to accelerate the learning process and increase the accuracy since the models are pre-trained with a huge database. For our covid-19 detection problem, we applied the transfer learning from CNN classification models: VGG-16, ResNet, Inception, Xception, DenseNet.

1. **VGG**: developed by Simonyan and Zisserman [36] and composed consists of 16 (for VGG-16) or 19 (for VGG-19) convolutional layers, it contains only 3x3

[1] IMAGENET. http://www.image-net.org/

convolutions, but a lot of filters. The final layers are two fully connected layers, each with 4,096 nodes, then followed by a softmax classifier. It is currently the most popular choice in the community for extracting features from images.

2. **ResNet**: introduced by Kaiming He et al. [37] and consists of several residual modules where each module represents a layer. It is a new architecture with skip connections applied to each layer of the network before the ReLu activation function, allowing to preserve the gradient. Using this technique, they have been able to form a neural network with 152 layers, but with lower complexity than VGGNet.
3. **Inception**: this architecture uses inception modules and aims to test all kind of convolution configuration to improve its performance by diversifying its attributes. It uses the 1×1 convolution to limit its computational complexity. The original version of this architecture was called GoogLeNet [38], but later manifestations were simply called Inception vN where N refers to the version number published by Google.
4. **Xception**: proposed by François Chollet [39], an extension of Inception's architecture that replaces Inception's standard modules with deeply separable convolutions.
5. **DenseNet**: the name DenseNet refers to Densely Connected Convolutional Networks. It was proposed by Gao Huang et al. in 2017 [40]. Traditional convolutional networks with n layers have n connections; one between each layer and its subsequent layer. DenseNet is composed of dense blocks, in these blocks, the layers are densely connected: each layer receives as input all the output characteristics of the previous layers.

3.1.4 models evaluation

Once the process of learning is completed, we can test the models with the test dataset which is not yet seen by the models. This allows to confirm the accuracy of our models and check the problems of overfitting.

3.2 Covid-19 detection using CT images

The use of CT scans allows to benefit from high resolutions and accurate sectional images or organs (lungs in this case). Within CT images, doctors can better differentiate between the types of fluids and thus provide an easier diagnosis for covid-19 detection. Moreover, physicians are also interested by the quantification of the size of Covid-19-related lesions, which is not possible with X-ray images. The dataset used for this study comes from the work of Zhao et al. [41], which contains 349 CT-Scan images from Covid-19, as well as 397 normal images. Actually, we are waiting a new and bigger CT-Scan dataset from CHU Ambroise Paré in Mons Belgium.
The detection of covid-19 using CT-scan represents a problem so similar to the one

seen in Section 3.1, where we need to classify images also, the only difference is present in the type of images (CT-scan instead of X-ray images). Thus, we applied the same classification Deep Learning architecture using CT-scan images.

4 Proposed approach for models explanation

This section is devoted to present our approach of Deep Learning explanation based on a combination of the related works and our best knowledge. Our explanation approach is based on four steps: problem identification, dataset collection, selection of Deep Learning models, explanation of Deep Learning models.

4.1 Problem identification

The explanation approach is always related to the type of application or problem, which is represented by images classification in our case. Before starting the process of explanation, it is so important to take in hand and understand the problem and the required solution. Several questions could be considered: what do we want to solve and why? Is there an interest of explainability? In fact, these questions may be considered before going further. This question is treated in Section 3 for our covid-19 detection problem.

4.2 Dataset collection

The problem resolution is highly dependent to the available data, which represent the most important element for solving a data science problem. The more you have, the less likely it is to have bias and overfitting problems. Finding a large amount of qualitative data representing a variety of situations is always necessary. The Visualization of data in various ways is also essential, to understand what is available and the results that will be obtained afterwards. This question is treated in Section 3.1.1 and 3.2 for our covid-19 detection problem.

4.3 Selection of Deep Learning models

The knowledge and selection of appropriate Deep Learning models is so important before starting explanation. As specified above, our focus is to solve a classification problem related to covid-19 detection using X-ray and CT-scan images. In this context, we have to select the best models in terms of accuracy and loss values after dividing data into training, validation and test sets. The general parameters will be defined, such as learning rate, loss, optimizer. All these parameters will be so useful for the explanation. This question is treated in Section 3.1.3 and 3.2 for our covid-19 detection problem.

4.4 Explanation of Deep Learning models

Once the problem, dataset and Deep Learning models are selected, we can start the process of analyzing and explaining neural networks and thus confirm their accuracy. In fact, the use of explainable artificial intelligence methods allows to ensure that the applied process is correct and close to the *Right to Explanation* requested. We propose an explanation based on six steps: XAI Frameworks identification and analysis, XAI Frameworks comparison and XAI methods selection, Visual comparison of XAI methods against the predicted class, Visual comparison of XAI methods against all classes, Non-visual evaluation of XAI methods, Model and XAI method selection.

4.4.1 XAI Frameworks identification and analysis

In literature, one can find several frameworks that use XAI methods, which mainly depend of the used Deep Learning framework (TensorFlow, Caffe, Pytorch, etc.). In this context, the XAI frameworks of *tf-explain*, *iNNvestigate*, *Skater*[2], *DeepExplain*[3] and *Deep Visualization Toolbox*[4] are compatible with tensorflow. On the other hand, *Captum* is compatible with PyTorch. Since we are working with Tensorflow for DL model development, we propose a brief description of tf-explain and iNNvestigate that are compatible with Tensorflow.

1. **tf-explain [42]:** is a Python library well adapted to Tensorflow 2.x. The proposed XAI methods are defined as a class, containing two main functions:
   - *explain*: where we can provide the model, the selected XAI method (defined in Section 2.1) and its parameters ;
   - *save*: allows to save the output of the *explain* function, as well as the path where to save the visualization of the explanation found.

2. **iNNvestigate [43]:** implements several methods of the state of the art with the both versions of Tensorflow (1.x and 2.x). Their objective is to simplify the analysis of neural networks

[2] Skater. https://github.com/oracle/Skater

[3] DeepExplain. https://github.com/marcoancona/DeepExplain

[4] Deep-visualization-toolbox. https://github.com/yosinski/deep-visualization-toolbox

4.4.2 XAI Frameworks comparison and XAI methods selection

The selected XAI frameworks (iNNvestigate and tf-explain) proposed several methods of explanation as shown in Table 1. Notice that the two XAI frameworks provide the implementation of three types of explanation methods: perturbation, gradient and relevance methods. In order to produce a complete explanation of methods, we start by selecting the XAI methods that are provided by the both libraries: Vanilla Gradient, Gradient * Input, SmoothGrad and Integrated Gradients. Thereafter, we will apply an analysis of the remaining methods in iNNversitgate framework and proxy models (LIME) provided from *Marco Tulio Correia Ribeiro* [5]. Notice that the last three methods in Table 1 : PatternNet & PatternAttribution [44] and DeepLift [45], implemented by iNNvestigate, do not provide convenient results for images classification and thus are not selected for this study. After this selection, we apply a comparison between the XAI methods in order to identify the most appropriate methods for our classification problem.

	tf-explain	**iNNvestigate**
Deconvolution		X
Activations Visualization	X	
Vanilla Gradient	X	X
Guided BackPropagation		X
Grad * Input	X	X
SmoothGrad	X	X
Integrated Gradients	X	X
LRP & Rules		X
Grad-CAM	X	
Occlusion	X	
DeepTaylor		X
PatternNet		X
PatternAttribution		X
DeepLift		X

Table 1: List of explanation methods implemented by *tf-explain* and *iNNvestigate*

4.4.3 Visual comparison of XAI methods against the predicted class

The results of explanation are so dependent to the provided parameters. In this section, we propose to compare visually the results where the idea to quantify the contribution of input image pixels to the results. First, we apply explanation using the class predicted by our model for the input image. i.e. the one with the highest probability. We visualize and interpret the results for this class. This approach is useful if the model presents a high score of classification accuracy.

[5] https://github.com/marcotcr/lime

4.4.4 Visual comparison of XAI methods against all classes

In case, where models can provide classification results with several candidate classes, it is important to analyze the results for the explanation of other classes, even if the probability is lower. Thus, for each existing class in the produced model, a comparison of the explanations is performed for several methods. This makes it possible to answer questions such as: "Why class A rather than class B?".

4.4.5 Non visual evaluation of XAI methods

In literature, the major methods of neural networks explanation provide a visual inspection of the result, which might be not sufficient in several situations mainly where the input images present high resolutions or where the target classes are present with very small sizes such as presented in a medical context. Therefore, we propose to offer a non-visual evaluation, inspired from the work of Samek et al. [46] based on the idea that perturbing important regions will have the most impact on the classification score. Taking into account that the saliency maps produce a decreasing ranking of the pixels related to their importance for the class score, a deletion of the most important pixels is made per step. This process is called *most relevant first*, abbreviated as MoRF. After each information removal, the effect and classification score are calculated. The evolution of the class score for different methods at each step (at each deleted pixel) is performed using this approach. This allows to evaluate the quality of the produced method.

4.4.6 Model and XAI method selection

The last step consists of comparing the DL classification models (related to our problem) in terms of accuracy and explainability. The idea is to determine the best compromise between the most complete explanation and accuracy. Based on this analysis, we can select the appropriate model and explanation method.

5 Experimental results

The experimental results are presented within three subsections where the first one presents the related results in terms of accuracy and loss for both X-ray and CT-scan covid-19 detection models. The second subsection is devoted to present the results of explaining the X-ray covid-19 detection models while the third subsection illustrates the results of explaining CT-scan covid-19 detection models. Notice the dataset is divided as follows: 70% for training, 15% for validation and 15% for test.

5.1 Covid-19 classification using X-ray and CT images

Table 2 presents the obtained values of loss and accuracy of our DL models (Section 3.1) for covid-19 detection using X-ray images. As shown in the table, the models present a high a accuracy ranging from 93% to more than 97%.

Model	Acc.	Loss
VGG16	0,9311	0,2239
VGG19	0,9361	0,1649
InceptionResnetV2	0,9741	0,1873
Inception v3	0,9568	0,2137
Xception	0,9435	0,1369
DenseNet	0,9601	0,1388
Resnet50	0,9686	0,1435

Table 2: Results of diffrent models using X-ray images

According to our model of covid-19 detection using CT scan images, the best result was obtained by applying a transfer learning from the VGG-16 architecture (where the weights are initiated using the ImageNet database). The obtained test accuracy is about 90% for classification, which can be considered as a good result in a medical context but we need to validate this accuracy using our explanation approach (Section 4). In fact, the validation of these models needs a deep evaluation of several questions:

- Among literature, which explanation methods can explain and interpret these models?
- How can we evaluate the accuracy of the selected explanation methods?
- Which models provide the best ratio accuracy/explainability?

These three questions are treated in the two next subsections.

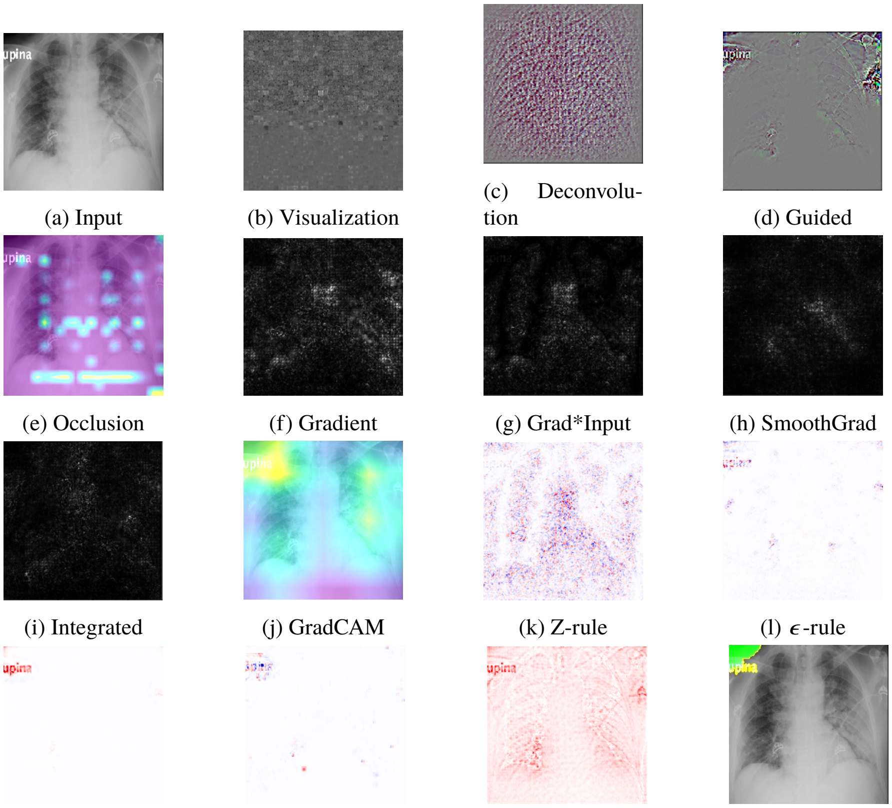
5.2 Explainable Covid-19 classification using X-ray

The explanation of our X-ray covid-19 classification models is applied with our approach proposed in Section 4. Several steps are followed:

5.2.1 Visual comparison of XAI methods against the predicted class

The first results are carried out on the DenseNet121 model using a Covid-19 image. With the deconvolution method (5c), the result is too noisy to define areas of interest. The occlusion method (5e) produces, in this case, inconsistent results with respect to the patch size parameter. Small variations in size produce large variations in results for any chosen size. The result obtained is therefore unreliable : https://www.youtube.com/watch?v=wzYAUFu0IYQ

The gradient methods (5f, 5g, 5h and 5i) give very similar results, which do not inform of the positive and negative contributions among the highlighted pixels. With the guided backpropagation (5d), the interest is mainly focused on two areas: top left where we find letters ("upina") and top right. This letter detection is also observable for the GradCAM method (5j), LIME (5p) and for DeepTaylor (5o) where the explanation proves to be much more intense on these letters than in the rest of the image. It is also part of the detection done for LRP rules, especially via LRP-PresetAFlat (5m), which only detects the letters in question. All this means that the main element contributing to the positive prediction of covid-19 is the detection of the letters present in the top left corner of the image.



(a) Input (b) Visualization (c) Deconvolution (d) Guided

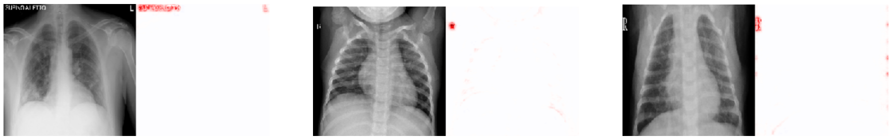
(e) Occlusion (f) Gradient (g) Grad*Input (h) SmoothGrad

(i) Integrated (j) GradCAM (k) Z-rule (l) ϵ-rule

(m) PresetAFlat (n) PresetBFlat (o) Deep Taylor (p) LIME

Fig. 5: DenseNet121 explanation with each method for a Covid-19 image

On Figure 6, we note that the letter detection is not an isolated case in the dataset. The model focus on parts of the image where there are letters (L,R), or it should logically focus on the lungs to distinguish the class of an image. Therefore, the conclusion to be drawn is that the patterns detected by the neural network are biased by the presence of the letters in the image.



(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

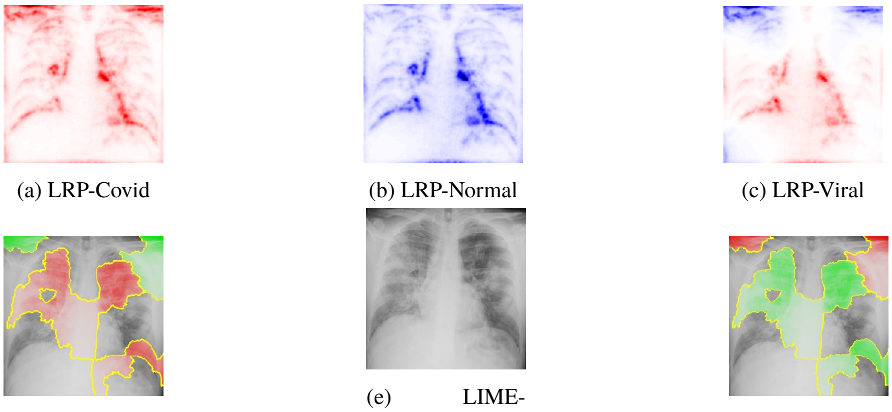
Fig. 6: LRP-PresetAFlat explanation for a Covid19, Normal, and Viral Pneumonia Image with DenseNet121

5.2.2 Visual comparison of XAI methods against all classes

In Figure 7 (image without letters in this case), we observe that the normal class is explained only by negative contributions by LRP, and by weights too low to be represented by the threshold defined with LIME (0.003), which is in agreement with the associated probability of 0.01%. Then, whether LRP or LIME is used, the viral pneumonia class offers the same perspective, predicting elements negatively at the level of the shoulder blades, and positively inside the lungs for a final probability of 2.2%. For the class predicted (Covid-19) with 97.7%, the scapulae are a strong element of prediction whether it is with LIME where they are the main positive weights, or with LRP where it is a more intense part of the prediction. This should not be the case and is another possible bias in the model.

5.2.3 Non visual evaluation of XAI methods

In Figure 8, the value by which the pixels are replaced is black (the minimum value). LRP-rules, Guided Backpropagation, Input*Gradient and Integrated Gradients get a high score. Notice that some methods that are not yet integrated in iNNvestigate are not tested, such as LIME and GradCAM which gave promising visual results, in line with the other methods.



(a) LRP-Covid

(b) LRP-Normal

(c) LRP-Viral

(d) LIME-Covid

(e) LIME-Normal

(f) LIME-Viral

Fig. 7: LRP-PresetAflat & LIME explanation for each class for a Covid-19 image

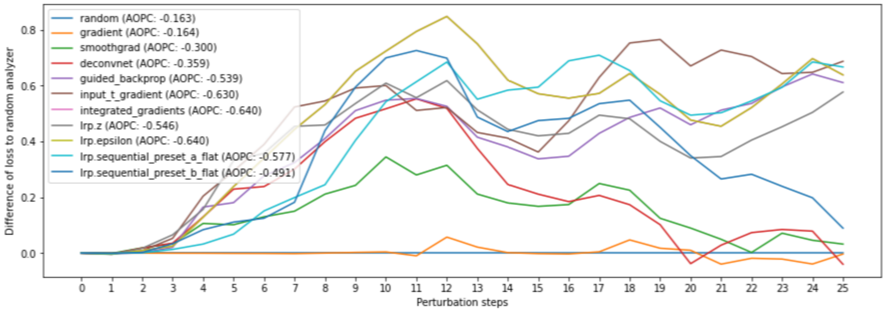
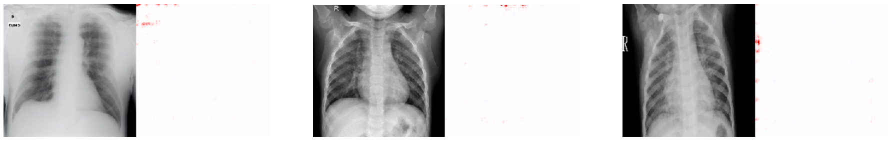


Fig. 8: Loss difference for different iNNvestigate methods compared to a random analyzer at each perturbation step, using the VGG-16 model applied to the Covid-19 dataset

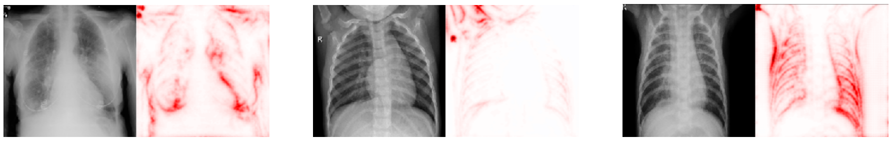
5.2.4 Model and XAI method selection

Following this detected bias, we applied our analysis on all available models in order to compare and confirm results. In this context, the VGG16 model, while clearly using letters for detection, seems to use other information in the images as seen in Figure 10. We note multiple red areas that are not related to the letters at all. These red areas are generally related to the lungs, suggesting a better model than conventionally expected. For the normal class, it is not only the letters that are considered. The attention is also focused on the top of the picture, at the head and shoulder level. By looking at the data (Figure 11), two elements can be noticed : the normal images are all taken higher than the others, *i.e.* we systematically see the jaw or at least the end of the neck, whereas for the other classes we never see the jaw or sometimes part of the shoulders. It is logical for such set of data to differentiate one normal class from another based on those elements. Furthermore, all the normal images and some viral pneumonia images have the characteristic of having the arms facing upwards. After discussion with a doctor, it turns out that the images of x-rays with arms facing upwards is a characteristic of x-rays taken with children. This can also be confirmed if one considers the humerus which is not fully developed.



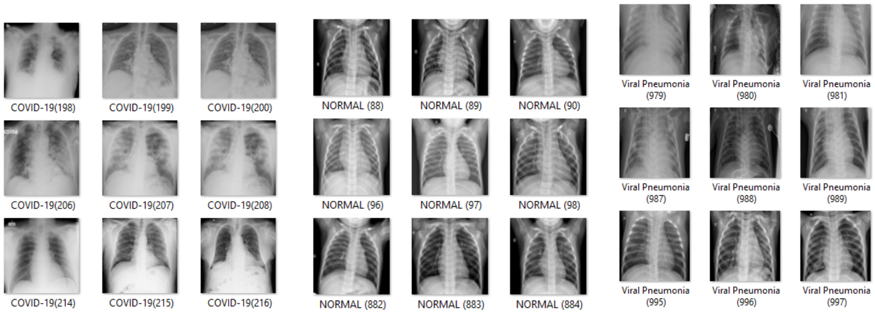
(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

Fig. 9: LRP-PresetAFlat explanation for a Covid19, Normal, and Viral Pneumonia Image with ResNet50



(a) Covid-19 Image (b) Normal Image (c) Viral Pneumonia Image

Fig. 10: LRP-PresetAFlat explanation for a Covid19, Normal, and Viral Pneumonia Image with VGG-16



(a) Covid19 Images (b) Normal Images (c) Viral Pneumonia Images

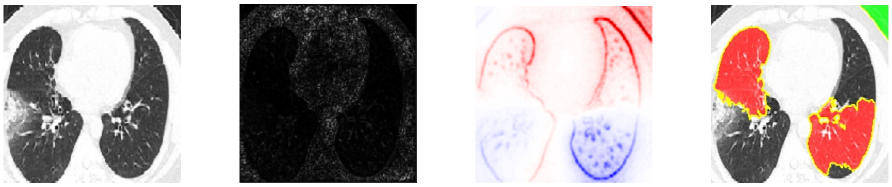
Fig. 11: Comparison between images from Covid-19, Normal, and Viral Pneumonia classes

5.2.5 Global Analysis

The methods of explicability have allowed to get an important observation: the identification of bias. Thanks to XAI, the defects of the dataset could be detected: for the normal class, arms mostly turned upwards, x-ray image taken higher than the other classes (showing the shoulder blades and the bottom of the head), and images only of children. Apart from that, for all classes, we find letters in the pictures for which the models give importance, when they should not. Therefore, we cannot rely on the models obtained. Among these, the explainability methods allowed to select the "best" model according to the plausibility of the explanations obtained (VGG-16 provides more plausible explanations, with interest not only in the letters but also in the lungs). Without the explicability methods, anyone would have stopped at the test score obtained by the model(s) and would have selected the highest precision, which proves that this is a very bad practice.

5.3 Explainable Covid-19 classification using CT images

In this part, only the methods giving the most interesting results are presented. In case of Covid-19 detection using CT images (Figure 12), the Integrated Gradients method has a majority of pixels outside the lungs, which is supposed to be the area of interest. In addition to this, LRP-PresetAFlat and LIME show the greatest interest in the upper right corner of the image, totally outside of what the CT-scan detects. This again represents a bias that should not exist, and the model is unreliable.
As a result of all these biases detected with x-ray as well as CT-scan images, classification alone is most probably not sufficient to correctly detect Covid-19 in the images with the available data sets. A new approach is brought to give to the neural network only the areas that are of interest: the lungs segmentation.



(a) Unsegmented Covid-19 image

(b) Integrated Gradients

(c) LRP PresetAFlat

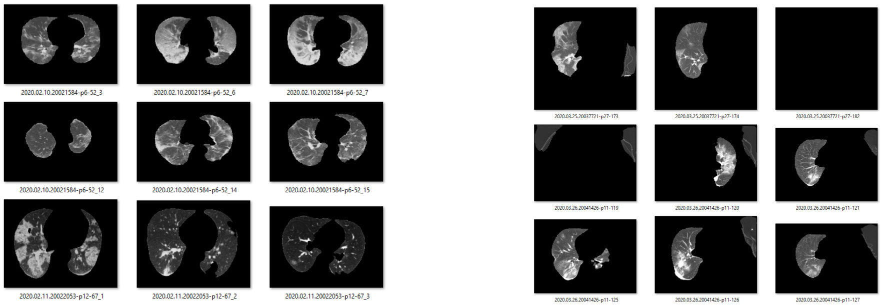
(d) LIME Proxy Model

Fig. 12: Unsegmented Covid-19 CT-Scan explained for a VGG-16 model

5.3.1 Preprocessing segmentation

To focus the interest of the network on the lungs, the other areas of the image must be removed. To do this, the lung segmentation model R231-Covidweb (U-Net) from Hofmanninger et al. [47] is used on all images in the dataset. The results are sorted in order to keep only the correct segmentations. This reduces the dataset to 233 Covid-19 images and 293 other images.

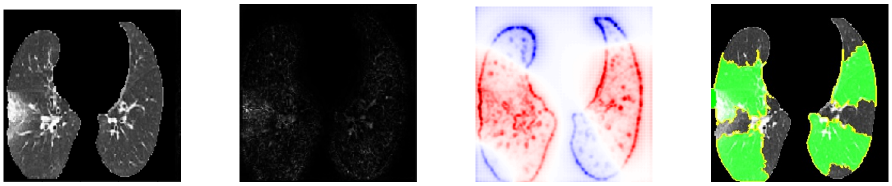
After training from new the VGG-16 model on the remaining images, the difference in result can be seen on Figure 14. The removal of uninteresting areas forces the network to focus on the remaining area, which consequently limits the biases learned.


2020.02.10.20021584-p6-52_3
2020.02.10.20021584-p6-52_6
2020.02.10.20021584-p6-52_7
2020.02.10.20021584-p6-52_12
2020.02.10.20021584-p6-52_14
2020.02.10.20021584-p6-52_15
2020.02.11.20022053-p12-67_1
2020.02.11.20022053-p12-67_2
2020.02.11.20022053-p12-67_3
2020.03.25.20037721-p27-173
2020.03.25.20037721-p27-174
2020.03.25.20037721-p27-182
2020.03.26.20041426-p11-119
2020.03.26.20041426-p11-120
2020.03.26.20041426-p11-121
2020.03.26.20041426-p11-125
2020.03.26.20041426-p11-126
2020.03.26.20041426-p11-127

(a) Successful segmentation

(b) Failed segmentation

Fig. 13: Visualization of good and bad results sorted from lungs segmentation



(a) Segmented Covid image

(b) Integrated Gradients

(c) LRP PresetAFlat

(d) LIME Proxy Model

Fig. 14: Segmented Covid-19 CT-Scan explained for a VGG-16 model

6 Conclusion

In this work, we addressed the problem of covid-19 diagnosis using chest X-ray and CT-Scan images. In order to confirm the ability of our models to differentiate covid-19 X-ray and CT-Scan images from both healthy persons and pneumonia patients, we performed a study on different deep learning models for the classification of covid-19 images. In this work, two public datasets were used, the X-ray dataset that contains 219 COVID-19 images, 1341 normal images, and 1345 viral pneumonia images. On the other side, the CT-Scan dataset contains 349 images from covid-19 and 397 normal images. The obtained results showed that the transfer learning of the models applied to the used datasets offers good performances. Moreover, we performed a large explainability analysis to interpret and visualize how our models work. Experimental results showed the interest of our explanation approach for the identification of the most interpretable DL model, the measure of the positive and negative contribution of input parameters in the decision of DL models, and the detection of data biases. The provided explanations were evaluated by doctors and physicians that confirmed the efficiency of our models. As future work, we plan to extend our dataset in order to train more accurate models and reduce biases data thanks to DL explanation.

References

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
2. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.
3. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
4. C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.
5. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
6. Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural network see the world-a survey of convolutional neural network visualization methods," *arXiv preprint arXiv:1804.11191*, 2018.
7. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.
8. J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
9. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

10. 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.
11. P.-J. Kindermans, K. Schütt, K.-R. Müller, and S. Dähne, “Investigating the influence of noise and distractors on the interpretation of neural networks,” *arXiv preprint arXiv:1611.07270*, 2016.
12. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.
13. D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” *arXiv preprint arXiv:1706.03825*, 2017.
14. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PloS one*, vol. 10, no. 7, 2015.
15. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, “Layer-wise relevance propagation: an overview,” in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.
16. G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
17. 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.
18. J. R. Zilke, E. L. Mencía, and F. Janssen, “Deepred–rule extraction from deep neural networks,” in *International Conference on Discovery Science*. Springer, 2016, pp. 457–473.
19. X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, J. Liu, K. Xu, L. Ruan, J. Sheng, Y. Qiu, W. Wu, T. Liang, and L. Li, “A deep learning system to screen novel coronavirus disease 2019 pneumonia,” *Engineering*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2095809920301636
20. M. Rahimzadeh and A. Attar, “A modified deep convolutional neural network for detecting covid-19 and pneumonia from chest x-ray images based on the concatenation of xception and resnet50v2,” *Informatics in Medicine Unlocked*, vol. 19, p. 100360, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352914820302537
21. N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, “Automated deep transfer learning-based approach for detection of covid-19 infection in chest x-rays,” *IRBM*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1959031820301172
22. Y. Pathak, P. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. Shukla, “Deep transfer learning based classification model for covid-19 disease,” *IRBM*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1959031820300993
23. L. Sarker, M. Islam, T. Hannan, and Z. Ahmed, “Covid-densenet: A deep learning architecture to detect covid-19 from chest radiology images,” *Preprints*, 2020. [Online]. Available: https://www.preprints.org/manuscript/202005.0151/v1
24. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” 2017.
25. 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 (ICCV)*, Oct 2017.
26. F. Ucar and D. Korkmaz, “Covidiagnosis-net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (covid-19) from x-ray images,” *Medical Hypotheses*, vol. 140, p. 109761, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306987720307702
27. T. Mahmud, M. A. Rahman, and S. A. Fattah, “Covxnet: A multi-dilation convolutional neural network for automatic covid-19 and other pneumonia detection

from chest x-ray images with transferable multi-receptive feature optimization," *Computers in Biology and Medicine*, vol. 122, p. 103869, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010482520302250

28. M. R. Karim, T. Dohmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, and O. Beyan, "Deepcovidexplainer: Explainable covid-19 diagnosis based on chest x-ray images," 2020.
29. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, mar 2018, pp. 839–847. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/WACV.2018.00097
30. B. K. Iwana, R. Kuroki, and S. Uchida, "Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation," 2019.
31. L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105608, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260720314413
32. M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi, M. B. I. Reaz, and T. I. Islam, "Can ai help in screening viral and covid-19 pneumonia?" 2020.
33. S. I. di Radiologia Medica e Interventistica, "Covid-19 database: Casistica radiologica italiana," 2020.
34. J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," 2020.
35. D. S. K. et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Engineering*, 2020. [Online]. Available: https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5
36. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
37. 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.
38. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
39. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," 2016, cite arxiv:1610.02357. [Online]. Available: http://arxiv.org/abs/1610.02357
40. 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.
41. J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," *arXiv preprint arXiv:2003.13865*, 2020.
42. "tf-explain," https://tf-explain.readthedocs.io/en/latest/, accessed: 2020-28-05.
43. M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "innvestigate neural networks," *Journal of Machine Learning Research*, vol. 20, no. 93, pp. 1–8, 2019.
44. P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," *arXiv preprint arXiv:1705.05598*, 2017.
45. A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3145–3153.
46. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
47. J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem," *arXiv preprint arXiv:2001.11767*, 2020.