

EXPLORATION OF INTERPRETABILITY TECHNIQUES FOR DEEP COVID-19 CLASSIFICATION USING CHEST X-RAY IMAGES

Soumick Chatterjee^{*1,2,4}
*Rupali Khatun*⁶

Fatima Saad^{*3,4}
Petia Radeva^{6,7}
Oliver Speck^{2,4,8,9,10}

Chompunuch Sarasaen^{*2,3,4}
Georg Rose^{3,4}
*Andreas Nürnberger*¹

Suhita Ghosh^{*5,4}
*Sebastian Stober*⁵

¹Data and Knowledge Engineering Group, Otto von Guericke University, Magdeburg, Germany

²Biomedical Magnetic Resonance, Otto von Guericke University, Magdeburg, Germany

³Institute for Medical Engineering, Otto von Guericke University, Magdeburg, Germany

⁴Research Campus STIMULATE, Otto von Guericke University, Magdeburg, Germany

⁵Artificial Intelligence Lab, Otto von Guericke University, Magdeburg, Germany

⁶Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain

⁷Computer Vision Center, Cerdanyola, Barcelona, Spain

⁸German Center for Neurodegenerative Diseases, Magdeburg, Germany

⁹Center for Behavioral Brain Sciences, Magdeburg, Germany

¹⁰Leibniz Institute for Neurobiology, Magdeburg, Germany

ABSTRACT

The outbreak of COVID-19 has shocked the entire world with its fairly rapid spread and has challenged different sectors. One of the most effective ways to limit its spread is the early and accurate diagnosis of infected patients. Medical imaging such as X-ray and Computed Tomography (CT) combined with the potential of Artificial Intelligence (AI) plays an essential role in supporting the medical staff in the diagnosis process. Thereby, five different deep learning models (ResNet18, ResNet34, InceptionV3, InceptionResNetV2, and DenseNet161) and their Ensemble have been used in this paper, to classify COVID-19, pneumoniae and healthy subjects using Chest X-Ray images. Multi-label classification was performed to predict multiple pathologies for each patient, if present. Foremost, the interpretability of each of the networks was thoroughly studied using techniques like occlusion, saliency, input X gradient, guided backpropagation, integrated gradients, and DeepLIFT. The mean Micro-F1 score of the models for COVID-19 classifications ranges from 0.66 to 0.875, and is 0.89 for the Ensemble of the network models. The qualitative results depicted the ResNets to be the most interpretable models.

Index Terms— COVID-19, Pneumonia, Chest X-ray, Multi-label Image Classification, Deep Learning, Model Ensemble, Interpretability Analysis

1. INTRODUCTION

In 2020, the world has witnessed a serious new global health crisis: the outbreak of the infectious COVID-19 disease which is

caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1, 2]. On March 11, 2020, COVID-19 was declared a global pandemic by the World Health Organization (WHO) due to the dramatically increasing number of infected people over multiple countries and continents¹. As of May 31, 2020, more than 5.9 million COVID-19 cases have been confirmed globally with a fatal rate of over 6.1%². COVID-19 has highly challenged the healthcare systems worldwide not to collapse mainly due to the shortage of medical supplies and staff.

Owing to the long incubation period of COVID-19 and its high contagiousness nature, it is important to identify the infected cases at an early stage and to isolate them from the healthy population, especially with the absence of vaccines and specific therapeutic protocols. So far, viral nucleic acid detection using Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been considered as the golden reference standard diagnostic method for COVID-19 cases³. However, it was reported that RT-PCR tests suffer from a high rate of false-negatives mainly due to laboratory and sample collection errors [3, 4]. In practice, this means that some COVID-19 patients may not be detected and given the right treatment which might lead to a widespread of the virus to other healthy subjects. Additionally, this diagnosis process is time-consuming as it takes more than four hours to

¹WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 available at <https://www.who.int/dg/speeches>

²WHO Situation Report-132 available at <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>

³Accessed on 31 May 2020 <https://radiopaedia.org/articles/covid-19-3>

*S. Chatterjee, F. Saad, C. Sarasaen, S. Ghosh contributed equally to this work.

receive the test results [5]. These limitations make the RT-PCR method unfavorable in clinical practice.

On the other hand, medical imaging arises as a tremendous alternative candidate for the screening of COVID-19 cases and for discriminating them from other conditions, whereas most of the COVID-19 patients show abnormalities in medical chest imaging [6, 7, 8]. In this context, Chest radiography (CXR) and Computed Tomography (CT) have been widely used in front-line hospitals for diagnosing COVID-19 cases [9, 10, 11]. In some cases, it was shown that chest CT images have exhibited higher sensitivity than RT-PCR and have detected COVID-19 infections in patients with negative RT-PCR results [11, 12]. Recent COVID-19 radiology literature has revolved around CT imaging primarily due to its higher sensitivity. However, there are several advantages of fostering the use of CXR imaging for the assessment of COVID-19 cases. X-ray imaging is cheaper, easier to perform, and less harmful than CT [13]. Moreover, X-ray machines are much more available than CT scanners, especially in developing countries. In addition, with the help of portable X-ray machines, imaging can be performed in the isolation rooms, decreasing the risk of infection transmission during transportation to the CT room, as well as the time needed for disinfecting the CT equipment and room [14].

Airspace Opacities or Ground-Glass Opacities (GGO) are the commonly reported radiological appearances with COVID-19. Bilateral, peripheral, and lower zone predominant distributions are mostly observed (90%) [15]. However, these manifestations are very similar to various viral pneumoniae and other inflammatory and infectious lung diseases. Hence, it is difficult for radiologists to discriminate COVID-19 from other types of pneumoniae [16]. Expert radiologists are needed to achieve high diagnostic performance and the diagnosis duration required is relatively high. In this context, it appears that Artificial Intelligence (AI) can play one of the potential roles in strengthening the power of the imaging tools for fighting COVID-19. AI technologies have been applied and integrated into the imaging workflow to support a contactless and automated patient posing and positioning [17]. Moreover, many AI applications have focused on infection quantification and identification in order to fully automate the diagnosis decision and help the medical specialists. This has played an essential role in accelerating the diagnosis for radiologists who are not specialized in COVID-19 diagnosis.

Several works on AI-assisted diagnosis were reported and promising results have been shown [18]. The classification of COVID-19 and other types of pneumonia has been investigated using deep learning techniques [6, 19]. However, to the best of our knowledge, most of these works lacked the discussion on the interpretability part: where did the networks focus to find the discriminative features. This leads to the hesitant use of deep learning techniques in clinical practice. With the "black box" settings and without the human verification, AI-based diagnosis is difficult as well as dangerous to be placed into practice. Thereby, in this work, the authors have considered the state-of-the-art deep learning models to classify COVID-19 and similar pathologies along with a thorough look into the interpretability of each of these models.

Foremost, motivated by the fact that one patient can have multiple pathologies at the same time, a multi-label classification was performed. The motivation behind considering deep learning and not the interpretable non-deep-learning techniques is mainly due to the fact that in recent times deep learning techniques has been observed outperforming the others in many challenges [20, 21, 22].

The remaining of the paper is organized as follows: in Section 2 most of the related works are reported and discussed, then in section 3, the strategy to create the dataset and the architecture design are exposed. Section 4 illustrates the classification results and the interpretability analysis. The results are analyzed in Section 5 and finally, Section 6 concludes the work and provides directions for further research.

2. RELATED WORKS

The spread of COVID-19 has attracted many researchers to concentrate their efforts towards developing AI-based approaches, for detection of the same from the various medical imaging modalities. Many efforts have been made to automate the diagnosis of COVID-19, by treating it as a multi-class classification task. [7] used ResNet50, InceptionV3 and InceptionResNetV2 models to classify COVID-19 patients using CXR images. They showed that the pre-trained ResNet50 model yields the highest accuracy (98%). However, they only discriminated between healthy and COVID-19 subjects but did not include the other types of pneumonia. Additionally, accuracy is considered to be a misleading metric in case of imbalanced datasets. [18] designed the COVID-Net for the detection of COVID-19 cases using CXR images. They used datasets including patients with bacterial pneumonia, viral pneumonia, COVID-19, and also healthy subjects. [6] used a ResNet-based model to classify between COVID-19 and non-COVID-19 patients. They achieved a sensitivity of 96% and a specificity of 70.7%. [23] presented a Dropweights based Bayesian Convolutional Neural Networks (BCNN) for CXR-based COVID-19 diagnosis. They showed interesting results regarding the estimation of the diagnosis decision uncertainty.

3. MATERIALS AND METHODS

3.1. Dataset

The CXR images were collected from two public datasets. The first dataset was the COVID-19 image data collection by [19]⁴, consisting of 236 images of COVID-19, 12 images of COVID-19 and ARDS, 4 images of ARDS, 1 image of Chlamydomphila, 1 image of Klebsiella, 2 images of Legionella, 12 images of Pneumocystis, 16 images of SARS, 13 images of Streptococcus and 5 images without any pathological findings. The second dataset was the Chest X-Ray Images (Pneumonia) dataset by [24]⁵, which

⁴Dataset available at: <https://github.com/ieee8023/covid-chestxray-dataset>

⁵Dataset available at: <https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>

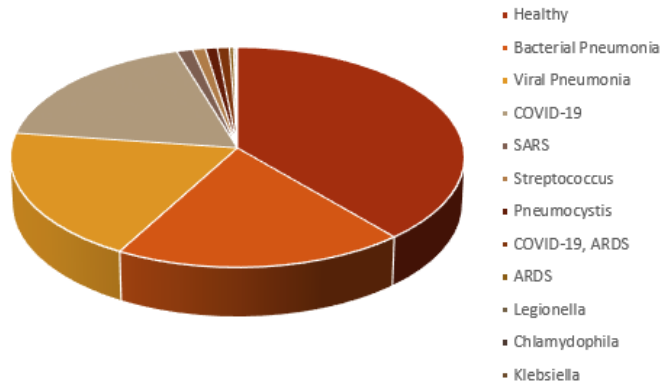


Fig. 1. CXR images distribution for each infection type in the dataset

has a total of 1583 images of healthy subjects, 1493 images of viral pneumonia and 2780 of bacterial pneumonia. From this dataset, 500 images of healthy, 250 images of viral pneumonia and 250 images of bacterial pneumonia, were randomly chosen. Fig 1 portrays the final data distribution considered for the work. This dataset of CXR images consists of posterior-anterior (PA), anterior-superior (AP), and anterior-superior supine (AP supine) radiographs. Although AP view is not the priority positioning and has disadvantages such as overlapping of organs which might interfere with the network prediction⁶, it is a technique commonly used for COVID-19 patients who are in coma.

The hierarchical nature of pathologies can be observed in this combined dataset. SARS and COVID-19 are sub-types of viral pneumonia. On the other hand, Streptococcus, Klebsiella, Chlamydomphila, and Legionella are sub-types of bacterial pneumonia, and Pneumocystis is a sub-type of fungal pneumonia. Furthermore, viral, bacterial, and fungal pneumoniae are different types of pneumonia. Therefore, a patient having COVID-19 is inherently having viral pneumonia. Additionally the dataset comprises cases where a patient has both COVID-19 and ARDS, which makes it suitable for multi-label classification.

The final dataset was randomly divided into a training set, consisting of 60% of the unique subjects and the remaining 40% of the subjects were used as a test set. 5-fold cross-validation (CV) was performed to evaluate the generalization capabilities of the models. The performance of the models during the 5-folds CV is reported in the sub-section 4.1. For interpretability analysis, only the results from the first fold were used, as it produced the best micro F1 scores.

3.2. Data Pre-processing

The dataset used for the task comprises X-ray images collected at different centers using different protocols and varying in size and

⁶Chest Radiograph <https://radiopaedia.org/articles/chest-radiograph?lang=us> Accessed: 2020-05-31

intensity. Therefore, all the images were initially pre-processed to have the same size. For making the image size uniform throughout the dataset, each of the images was interpolated using bicubic interpolation, to have 512 pixels on the longer side. The number of pixels in the shorter side was calculated preserving the aspect ratio of the original image. After that, zero-padding was used on the shorter side to make that side having 512 pixels, resulting in a 512 x 512 image. Image resizing was followed by percentile cropping, where the image intensity was cropped to 1st and 95th percentile, and then intensity normalisation to the interval [0,1] was performed. The percentile cropping normalisation minimizes the effect of intensity variation due to the non-biological factors.

3.3. Network models

During the course of this research, various network architectures were explored and experimented with, including several variants of VGG [25], ResNet [26], ResNeXt [27], WideResNet [28], Inception [29], DenseNet [30]. Prior to training on the dataset of this research work, all the networks were initialized with weights pre-trained on ImageNet. After observing the results, 5 network architectures were shortlisted for further analysis and were also used to create an Ensemble for better prediction performance. The models were selected based on different criteria, such as performance, the complexity of the model, etc. The selected models are discussed in this section and Table 1 shows the complexity of the models.

3.3.1. ResNet

At the nascent stage of deep learning, the deeper networks faced the problem of vanishing gradients/ exploding gradients [31, 32] which hampered the convergence. The deeper network faced another obstacle called degradation, where the accuracy starts saturating and degrading rapidly after a certain network depth. To overcome these problems, [26] designed a new network model called residual network or ResNet, where the authors came up with 'Skip Connection' identity mapping. This does not involve adding an extra hyper-parameter or learnable parameter, but just adding the output of the previous layer to the following layer. It unleashed the possibility of training deeper models without encountering the aforementioned problems.

After comparing against various versions of Resnet, two different variants, ResNet18, and ResNet34, were chosen for further analysis during this research.

3.3.2. InceptionNet

An image can have thousands of salient features. In different images, the focused features can be at any different part of the image making the choice of the right kernel size for a convolution network a very difficult task. A large kernel will have more focus on globally distributed information, while a smaller kernel will have a focus on local information. To overcome this problem, [29] came up with a new network architecture called InceptionNet or GoogleNet. The authors used filters of multiple sizes to operate on

the same level, which makes the network more "wider" rather than "deeper". To make it computationally more cost-effective, the authors limited the number of input channels by adding an extra 1x1 convolution before the 3x3 and 5x5 convolutions. Adding 1x1 convolutions is much cheaper than adding 5x5 convolutions. The authors introduced two auxiliary classifiers to prevent the problem of vanishing gradient, and an auxiliary loss is calculated on each of them. The total loss function is a weighted sum of the auxiliary loss and the real loss.

Too much reduction of dimensions can cause loss of information, also known as "representational bottleneck". To overcome this problem and to scale up the network in ways that it utilizes the added computation as efficiently as possible, the authors of InceptionNet introduced a new idea in [33] factorizing convolutions and aggressive regularization. The authors factorized each 5x5 convolution to two 3x3 convolution operations to improve computational speed. Furthermore, they factorized convolutions of filter size $n \times n$ to a combination of $1 \times n$ and $n \times 1$ convolutions. This network is known as InceptionV2.

In [33] the authors have also proposed InceptionV3, which extends InceptionV2 further by factorizing 7x7 convolutions, by label smoothing, and by adding BatchNorm in the auxiliary classifiers. Label smoothing is a type of regularizing component added to the loss formula that prevents the network from becoming too confident about a class.

InceptionV3 ranked in one of the top-5 positions during the initial trials and therefore was used for further analysis.

3.3.3. InceptionResNetV2

The different variants of InceptionNet and ResNet have shown very good performance with relatively low computational cost. With the hypothesis that residual connections would cause the training of Inception networks accelerated significantly, the authors of the original InceptionNet proposed InceptionResNet in [34]. In this, the pooling operation inside the main inception modules was replaced by the residual connections. Each Inception block is followed by a filter expansion layer (1x1 convolution without activation) which is used for scaling up the dimensions of the filters back before the residual addition, to match the input size.

This is one of the networks that has been used in this research, because of its performance on the dataset that has been used.

3.3.4. DenseNet

[30] came up with a very simple architecture to ensure maximum information flow between layers in the network. By matching feature map size throughout the network, they connected all the layers directly to all of their subsequent layers - a densely connected neural network, or simply known as DenseNet. DenseNet improved the information flow between layers by proposing this different connectivity pattern. Unlike many other networks like ResNet, DenseNets do not sum the output feature maps of the layer with the incoming feature maps but concatenates them.

During the initial trials of this work, DenseNet161 came up as a winner in terms of performance. So, in this research DenseNet161 was included.

Table 1. Number of trainable parameters in each model

<i>Model</i>	<i>No of parameters</i>
ResNet18	11,183,694
ResNet34	21,291,854
InceptionV3	24,382,716
DenseNet161	26,502,926
InceptionResNetV2	54,327,982

3.4. Interpretability techniques

Interpretability techniques can help understand the reasoning of a network for its predictions. There are various techniques already in existence. Some of the methods such as, Occlusion, Saliency, Input X Gradient, Integrated Gradients, Guided Backpropagation, DeepLIFT, which were explored in this research paper are explained briefly in this section. Apart from these methods, other model attribution methods like Guided GradCAM [35], Feature Ablation [36], Shapley Value Sampling [37], and layer attribution methods like Layer Conductance [38], Internal Influence [39] and many other methods have also been implemented and are part of the developed interpretability pipeline, but have not been explored during the course of this research.

3.4.1. Occlusion

Occlusion is one of the simplest interpretability techniques for image classifications. This technique helps to understand which features of the image steer the network towards a particular prediction or which are the most important parts for the network to classify a certain image. To get this answer, [40] performed an occlusion technique by systematically blocking different portions of the input image with a grey square box and monitoring the output of the classifier. The grey square is applied to the image in a sliding window manner, that moves across the image, obtaining many images, and then they are fed into the trained network to obtain probability scores for a given class for each mask position.

3.4.2. Saliency

In the context of visualisation, saliency refers to a topological representation of the unique features of an image. Saliency is one of the baseline approaches for the interpretation of the deep learning models. The saliency method of [41] returns the gradients of a model for its respective inputs. The positive values present in the gradients show that how a small change in the input image changes the prediction.

3.4.3. Input X Gradient

Input X Gradient is an extension of the Saliency approach. Similar to the saliency method of [41], this method of [42] also takes the gradients of the output with respect to the input, but additionally, multiplies the gradients by the input feature values.

3.4.4. Guided Backpropagation

Guided Backpropagation, also known as guided saliency, is another visualisation technique for deep learning classifiers. Guided backpropagation is a combination of vanilla backpropagation and deconvolution networks (DeConvNet) [43]. In this method, only the positive error signals get backpropagated and the negative signals are set to zero while backpropagating through a ReLU unit [44].

3.4.5. Integrated Gradients

[45] proposed a model interpretability technique, which assigns an importance score to each of the features of the input, by approximating the integral of gradients of the output for that input, along the path from the given references for the input.

3.4.6. DeepLIFT

Deep Learning Important FeaTures or DeepLIFT proposed by [46], is a method for pixel-wise decomposing the output prediction of a neural network on a specific input. This is done by backpropagating the contributions of all neurons in the network to every feature of the input. DeepLIFT compares the activation of each neuron to its reference activation, and then assigns contribution scores based on the difference. DeepLIFT can also reveal dependencies which might be missed by other approaches, by optionally assigning separate considerations to positive and negative contributions. Unlike other gradient-based methods, it uses difference from reference, which permits DeepLIFT to propagate an importance signal even in situations like where the gradient is set to zero.

3.5. Implementation setup

The models were implemented using PyTorch⁷. An interpretability pipeline for PyTorch-based classification models was developed with the help of Captum⁸ and PyTorch CNN Visualisations repository by [47]⁹. The code of this project: Diagno++¹⁰ and the interpretability pipeline: TorchEsegeta¹¹ are both available on GitHub.

Trainings were performed using Nvidia GeForce 1080 Ti and 2080 Ti GPUs, having 11GB of memory each. The loss was calculated using B Cross-Entropy (BCE) with Logits¹², which

combines Sigmoid layer with the BCE loss, to achieve better numerical stability than using the Sigmoid layer followed by BCE loss separately. The numerical stability is achieved by using the log-sum-exp trick, which can prevent underflow/overflow errors. The loss was minimized by optimizing the model parameters using the Adam optimizer [48], with a learning rate of 0.001 and a weight decay of 0.0001. A manual seed was used to ensure reproducibility¹³ of the models. Automatic Mixed Precision was used using Apex¹⁴, to speed-up the training and to decrease the GPU memory requirements.

The interpretability pipeline was used on the models using Nvidia Tesla V100 GPUs, having 32GB memory each. Some of the interpretability techniques could not be used on certain models, due to the lack of GPU memory caused by the complexities of the models.

In this multi-label classification setup, the model was trained to predict the disease and also its super-types. Hence, when a network encounters an image of a COVID-19 patient, it should ideally predict it as pneumonia, viral pneumonia, and COVID-19. When a network encounters an image of a patient having multiple pathologies, like in this dataset some patients have both COVID-19 and ARDS, ideally the network should predict it as pneumonia, viral pneumonia, COVID-19 as well as ARDS. Interpretability analysis was performed for each label of each image in the test set.

4. RESULTS AND INTERPRETABILITY ANALYSIS

In a multi-class setting, classifiers are generally evaluated with respect to precision, recall, and F1 metrics. In a multi-label classification setting, the same metrics are calculated in two ways: macro and micro averaging [49].

$$Macro = \frac{1}{P} \sum_{i=1}^P Metric \left(TP_i, FP_i, TN_i, FN_i \right) \quad (1)$$

As shown in Eq 1, the macro-based metrics are first computed individually from the true-positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of each class/pathology and then averaged, where P denotes the number of classes and $Metric \in \{\text{precision, recall, F1}\}$.

This manner of computation of the metrics causes to treat each pathology equally and the metric values get heavily influenced by the rare labels.

$$Micro = Metric \left(\sum_{i=1}^P TP_i, \sum_{i=1}^P FP_i, \sum_{i=1}^P TN_i, \sum_{i=1}^P FN_i \right) \quad (2)$$

In micro-based metrics, TP, TN, FP, and FN of each class/pathology are added individually and then averaged, as shown in Eq 2. Therefore, the micro-based metrics portray the aggregated contribution of all classes/pathologies. Therefore, the influence of the

⁷<https://pytorch.org/>

⁸<https://captum.ai/>

⁹<https://github.com/utkuozbulak/pytorch-cnn-visualizations>

¹⁰<https://github.com/soumickmj/diagnoPP>

¹¹<https://github.com/soumickmj/TorchEsegeta>

¹²<https://pytcs/stable/nn.html>

¹³<https://pytorch.org/docs/stable/notes/randomness.html>

¹⁴<https://github.com/NVIDIA/apex>

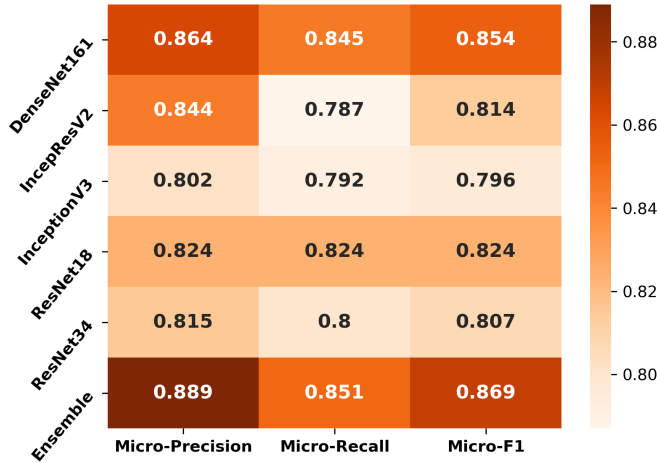


Fig. 2. Comparison of the classifiers based on micro metrics

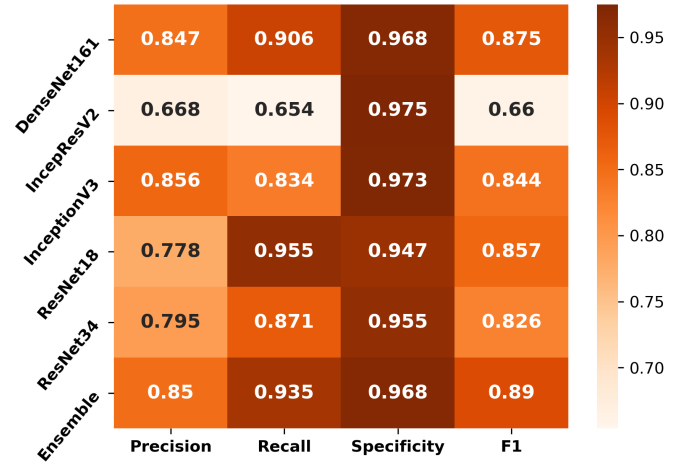


Fig. 3. Performance of the classifiers for COVID-19.

predictions out of the minority classes gets diluted among the contributions from the majority classes. This makes the micro-based metrics an appropriate measure to estimate the overall performance of the classifier, especially in case of imbalanced datasets. Since the dataset used was highly imbalanced, micro-based metrics have been considered for the evaluation of the classifiers [50].

4.1. Model outcome

4.1.1. Overall comparisons of the classifiers

Fig 2 portrays that the overall performance of the classifiers over pathologies was similar. Among the non-Ensemble models, DenseNet161 performed the best concerning all metrics. Although InceptionResNetV2 was the most complex model among all, it produced the worst recall, which implies the ability of the model to find the pathology affected cases was poor compared to the less complex models. ResNet18 was the least complex model among the non-Ensemble classifiers and it ranked second after DenseNet161 with respect to micro F1. The Ensemble produced the best results and minimum variance as portrayed in Table 2 over the 5-fold cross-validation.

4.1.2. Comparisons of the classifiers for different pathologies

The authors have also compared the classifiers' performance at the pathology level. The average metric values across 5 cross-validation folds have been depicted in Fig 3 to Fig 7 for COVID-19, pneumonia, viral pneumonia, bacterial pneumonia and healthy subjects respectively. While comparing the models using average F1, it has been observed that the performance of most of the models for COVID-19, pneumonia, and healthy was good, except for the performance of the InceptionResNetV2 for COVID-19 cases. Among all the models, the results of DenseNet161 were the most promising ones, for all the diseases. For COVID-19 classification, DenseNet161 performed the best,

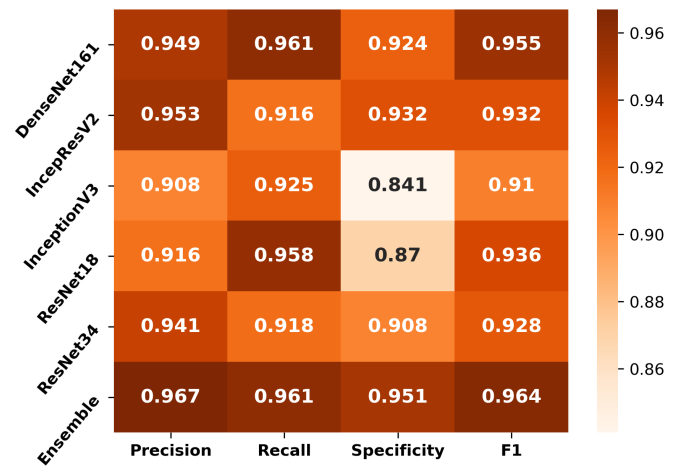


Fig. 4. Performance of the classifiers for Pneumonia

and ResNet18 has bagged the second position. DenseNet161 performed the best for pneumonia. InceptionResNetV2 provided the highest performance for viral pneumonia classification. Lastly, InceptionV3 gave the highest scores for bacterial pneumonia.

4.2. Interpretability of models

In the first sub-subsection 4.2.1 different interpretability techniques have been explored for different classifiers with respect to the different diseases. The second sub-section 4.2.2 talks about how the different models performed for specific pathologies.

All the given interpretability analyses were performed for that specific input CXR image which has been shown as the underlay.

Table 2. Performance of all the classifiers with respect to micro based metrics over 5-folds

<i>Model</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
DenseNet161	0.864 ± 0.012	0.845 ± 0.015	0.854 ± 0.008
InceptionResNetV2	0.844 ± 0.023	0.787 ± 0.063	0.814 ± 0.042
InceptionV3	0.802 ± 0.065	0.792 ± 0.044	0.796 ± 0.053
ResNet18	0.824 ± 0.014	0.824 ± 0.008	0.824 ± 0.007
ResNet34	0.815 ± 0.022	0.800 ± 0.025	0.807 ± 0.018
Ensemble	0.889 ± 0.010	0.851 ± 0.005	0.869 ± 0.007

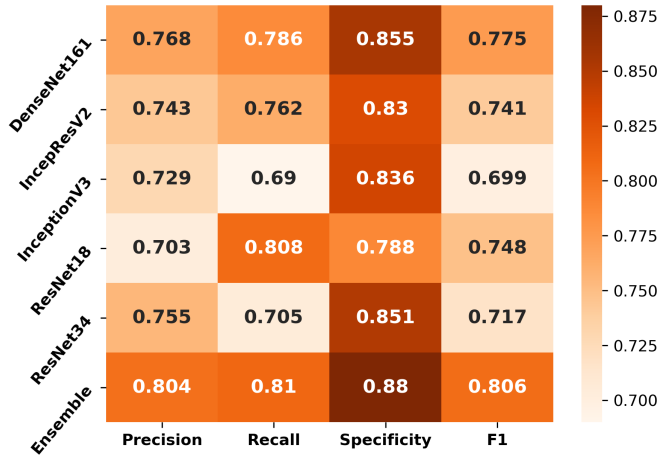


Fig. 5. Performance of the classifiers for Viral Pneumonia

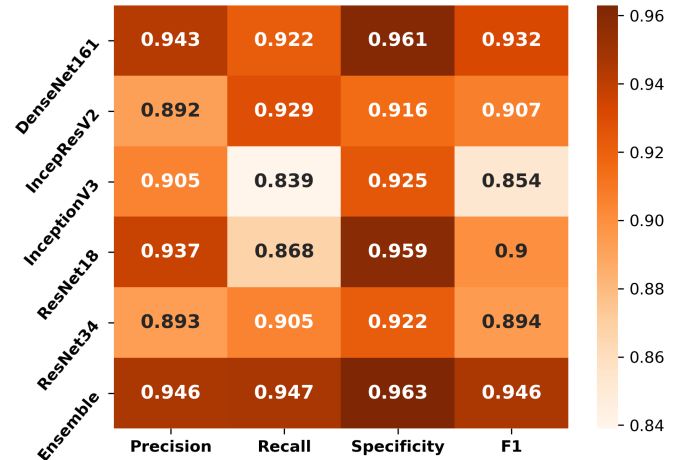


Fig. 7. Performance of the classifiers for Healthy subjects



Fig. 6. Performance of the classifiers for Bacterial Pneumonia

4.2.1. Pathology based comparisons of interpretability techniques for the models

To visualize the results on specific case, the models were interpreted using occlusion, saliency, inputXgradient, guided backpropagation and integrated gradients, and have been shown

in Fig 8, Fig 9 and Fig 10. Apart from occlusion, the other interpretability techniques failed to run for DenseNet161 due to GPU memory limitations. For DeepLIFT, ResNets encountered an additional problem because of the in-place ReLU operations used in those models. The models have to be updated to be able to run DeepLIFT on them.

According to the clinical findings of the COVID-19 image data provided by [19], multiple abnormalities of the lungs were located on the right upper and lower pulmonary field, as well as the upper left part of the lung. The models predicted this case as COVID-19, pneumonia, and viral pneumonia responding to the pathology of lung infection. One can see that the models' focus area for COVID-19 differ from the focus area for pneumonia and viral pneumonia. DenseNet161 and InceptionResNetV2 emphasized mostly on the right lung. InceptionV3, ResNet18, and ResNet34 covered both right and left parts, not only the lesion but also the irrelevant regions out of the lung.

4.2.2. Intense Interpretability

The failure case of the best performing model for COVID-19 classification: Even though the DenseNet161 performed the best among all the models, it gave false negative for some of the COVID-19 patients, whereas the rest of the models including the

Ensemble could correctly predict. The occlusion results of the models can be observed in Fig 11. The image is of a 70-year-old woman, who had three days of cough, myalgias, and fever; without any recent overseas travel. A series of chest X-ray images were obtained before the confirmation of coronavirus infection and the follow-ups were done in 3 days, 7 days, and 9 days. It shows the progression of radiographic changes.

On the other hand, ResNet18 is the most outstanding model, as it yielded high evaluation scores, despite having the least number of network parameters. The interpretability analysis of this model showed where the lesion was located and also the network can be utilized for the follow-up or severity estimations as illustrated in Fig 12.

COVID-19, pneumonia and viral pneumonia: Based on the fact that COVID-19 is a subset of viral pneumonia, the focus of this section is centralized on the interpretability comparison of the models for these three pathologies. The interpretability techniques reported that different networks focused on different areas for the same CXR image for predicting each of the diseases. It was observed that the focus area of DenseNet161 for COVID-19 was explicitly different from the one for pneumonia and viral pneumonia. However, InceptionResNetV2 and InceptionV3 emphasized on a similar area (different focus areas for each model) for all three pathologies. Furthermore, ResNet18 and ResNet34 targeted the lung region for COVID-19 and viral pneumonia but differed for pneumonia. Fig 13 exhibits the mentioned findings.

5. DISCUSSION

The literature review portrayed that the diagnosis of COVID-19 was seen as a multi-class classification task rather than a multi-label classification. The datasets used in the previous works vary in terms of the amount of data used for the classification task. [7] created a balanced dataset by appending the 50 COVID cases with 50 healthy cases from another dataset and reported the highest mean specificity score of 0.90 using InceptionV3. The others [6, 8, 18] performed a multi-class classification task on different imbalanced datasets using X-rays, and achieved the maximum mean specificity of 0.989, 0.979, 0.971 respectively. In this work, the InceptionResNetV2 achieved the highest specificity of 0.975, comparable to the previous works. However, in this research the authors have used a different dataset, train-test split, pre-processing techniques, compared to the previous works, which makes it unfair to compare the results with the previous works.

It is noteworthy that in some cases the network predicted the findings as a presence of COVID-19, while the radiologist (in the dataset label) did not report any abnormalities. It could be useful to have cooperation with radiologists to confirm these kinds of findings of the models. This also could imply that second opinion for diagnosis might be needed in such cases.

There were a couple of cases where the network detected both viral and bacterial pneumonia. According to [51] and [52], the induction of viral infection could lead to secondary bacterial infection and increase the severity of the symptoms. Though

such cases were considered as miss-predictions for the current dataset based on the available labels, one could argue that the network was able to detect such instances. These findings have to be confirmed by radiologists.

The main motivation to perform a multi-label classification over a multi-class classification was to be able to predict multiple pathologies from the images, if they are present. It was observed that all the networks, including the Ensemble, were able to correctly predict both COVID-19 and ARDS for the images which had both the pathologies present.

Lastly, this study also showed that the models could classify the lung pathologies from CXR images, although undesired objects, such as annotations or labels were obscuring the radiographs.

6. CONCLUSION AND FUTURE WORKS

In this paper, various deep learning based classifiers for multi-label classification of COVID-19 and similar pathologies in CXR images have been compared and the interpretability of those models has been investigated. In general, most of the models performed well. But, some of the models failed to perform on certain tasks. The authors have also created an Ensemble, which helped to fill-in those shortcomings of the models by combining their predictions. Moreover, it was observed that the smallest model ResNet18 competed well against considerably larger models. In fact, for certain situations, it performed better than the largest model in the mix, InceptionResNetV2. For patients who had more than one pathology, this multi-label classification setup was able to correctly predict all of those pathologies.

DenseNet161 was the best performing model in this setup, though it was observed that the focus of the network was many times on unrelated regions. After qualitative analysis, it can be said that the ResNets were the most interpretable models as the focus area of the networks were mostly on the correct regions.

Interpretability results obtained during this research will be investigated by radiologists to better understand the networks' reasonings from a clinical perspective. Certain miss-predictions of the networks can actually be errors in the dataset labels, which might be pointed out by a radiologist while thoroughly investigating the interpretability results.

There are various interpretability techniques for deep learning used for non-image data. For example, [53] talks about visualisation for speech recognition. It will be interesting to explore those methods on an image classification task, such as this one. Model explainability methods like LIME [54], SHAP [55] etc. have not been explored during this research but planned as future work.

This same approach can also be tried on CT images, to compare the networks' sensitivity for COVID-19 on CT and CXR images. Moreover, it would be interesting to investigate how the networks' performances are affected if completely unrelated pathologies (like tumours) are mixed with this current dataset.

Prior non-image information (like the patient's prior medical history, the result of RT-PCR test, etc.) can also be tried to be incorporated in the network models, to aid the networks in decision

making. Furthermore, instead of supplying the whole image to the models, a lung segmentation can be used as a pre-processing step, which might improve the networks' predictions by helping them to focus just on the region of interest which in this case are the lungs.

Training techniques like few-shot learning (including one-shot learning), semi-supervised learning, etc. can be explored for learning to classify COVID-19 cases from a small dataset. Moreover, joint segmentation-classification techniques can also be investigated for this multi-label classification problem. Several interpretation techniques are implemented in the interpretability pipeline, but have not been investigated in this paper, will be explored in the future for this dataset-model setup.

Acknowledgments

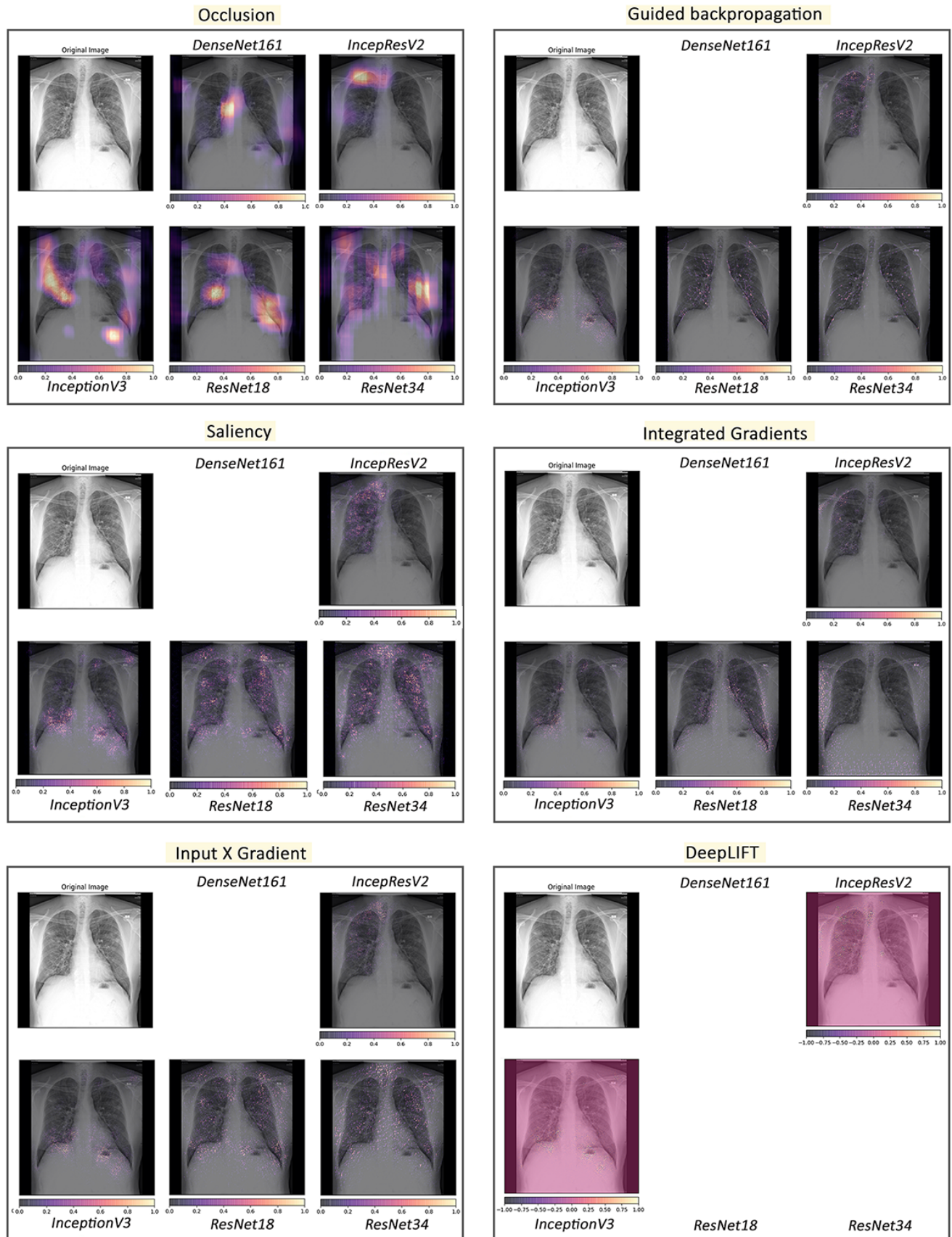
This work was conducted within the context of the International Graduate School MEMoRIAL at Otto von Guericke University (OVGU) Magdeburg, Germany, kindly supported by the European Structural and Investment Funds (ESF) under the programme Sachsen-Anhalt WISSENSCHAFT Internationalisierung (project no. ZS/2016/08/80646) and was partly funded by the Federal Ministry of Education and Research within the Forschungscampus STIMULATE under grant number 13GW0095A.

7. REFERENCES

- [1] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al., "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, 2020.
- [2] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al., "Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia," *New England Journal of Medicine*, 2020.
- [3] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, and Liming Xia, "Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases," *Radiology*, p. 200642, 2020.
- [4] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji, "Sensitivity of chest ct for covid-19: comparison to rt-pcr," *Radiology*, p. 200432, 2020.
- [5] Joungha Won, Solji Lee, Myungsun Park, Tai Young Kim, Mingu Gordon Park, Byung Yoon Choi, Dongwan Kim, Hyeshik Chang, V Narry Kim, and C Justin Lee, "Development of a laboratory-safe and low-cost detection protocol for sars-cov-2 of the coronavirus disease 2019 (covid-19)," 2020.
- [6] Jianpeng Zhang, Yutong Xie, Yi Li, Chunhua Shen, and Yong Xia, "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv preprint arXiv:2003.12338*, 2020.
- [7] Ali Narin, Ceren Kaya, and Ziyinet Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.
- [8] Ioannis D Apostolopoulos and Tzani A Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.
- [9] Jeffrey P Kanne, "Chest ct findings in 2019 novel coronavirus (2019-ncov) infections from wuhan, china: key points for the radiologist," 2020.
- [10] Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, et al., "Chest ct findings in coronavirus disease-19 (covid-19): relationship to duration of infection," *Radiology*, p. 200463, 2020.

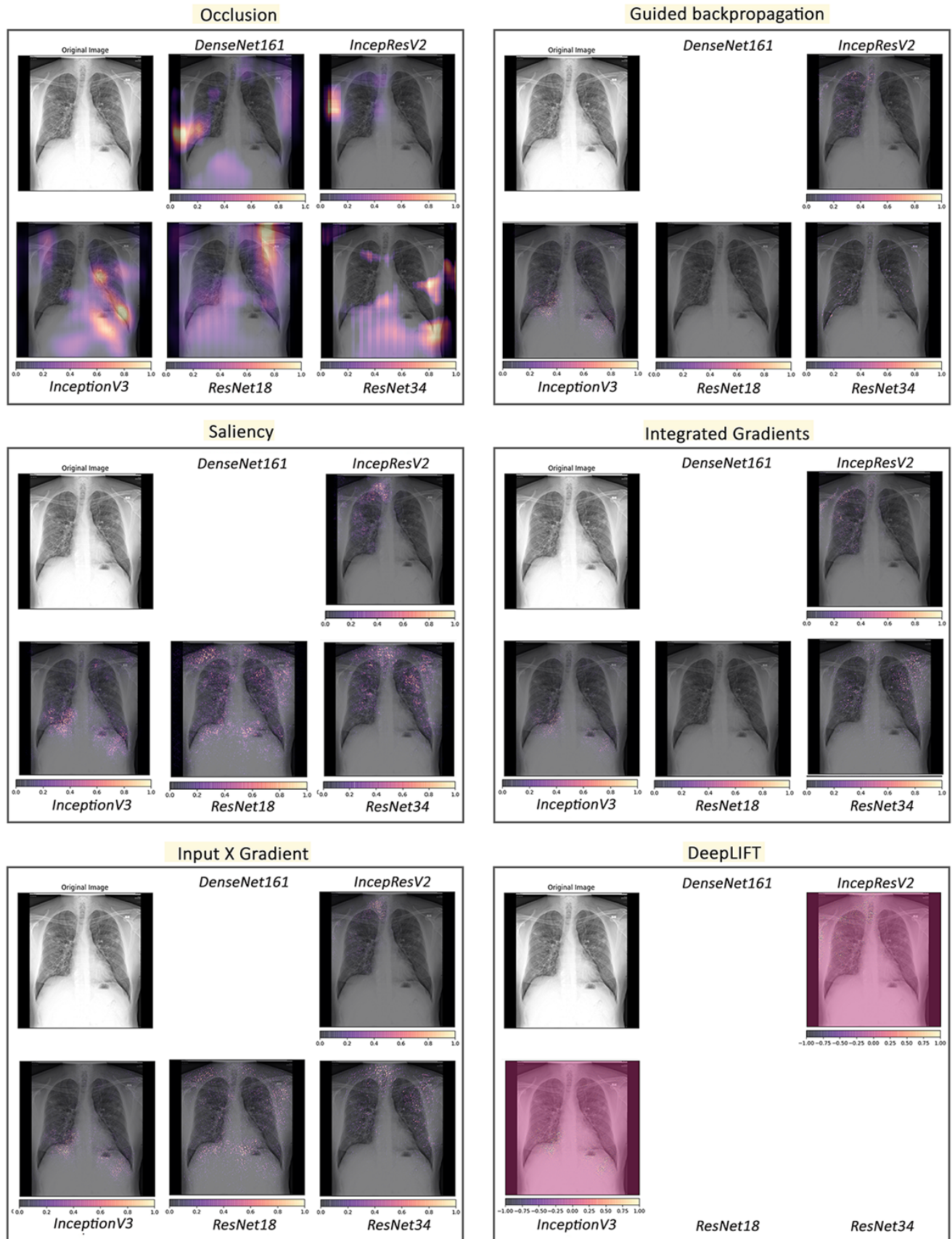
- [11] Xingzhi Xie, Zheng Zhong, Wei Zhao, Chao Zheng, Fei Wang, and Jun Liu, "Chest ct for typical 2019-ncov pneumonia: relationship to negative rt-pcr testing," *Radiology*, p. 200343, 2020.
- [12] Peikai Huang, Tianzhu Liu, Lesheng Huang, Hailong Liu, Ming Lei, Wangdong Xu, Xiaolu Hu, Jun Chen, and Bo Liu, "Use of chest ct in combination with negative rt-pcr assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, vol. 295, no. 1, pp. 22–23, 2020.
- [13] Geoffrey D Rubin, Christopher J Ryerson, Linda B Haramati, Nicola Sverzellati, Jeffrey P Kanne, Suhail Raoof, Neil W Schluger, Annalisa Volpi, Jae-Joon Yim, Ian BK Martin, et al., "The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society," *Chest*, 2020.
- [14] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber, "Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review," *Clinical Imaging*, 2020.
- [15] Ho Yuen Frank Wong, Hiu Yin Sonia Lam, Ambrose Ho-Tung Fong, Siu Ting Leung, Thomas Wing-Yan Chin, Christine Shing Yen Lo, Macy Mei-Sze Lui, Jonan Chun Yin Lee, Keith Wan-Hang Chiu, Tom Chung, et al., "Frequency and distribution of chest radiographic findings in covid-19 positive patients," *Radiology*, p. 201160, 2020.
- [16] Ming-Yen Ng, Elaine YP Lee, Jin Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, et al., "Imaging profile of the covid-19 infection: radiologic findings and literature review," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, pp. e200034, 2020.
- [17] Yang Wang, Xiaofan Lu, Yingwei Zhang, Xin Zhang, Kun Wang, Jiani Liu, Xin Li, Renfang Hu, Xiaolin Meng, Shidan Dou, et al., "Precise pulmonary scanning and reducing medical radiation exposure by developing a clinically applicable intelligent ct system: Toward improving patient care," *EBioMedicine*, vol. 54, pp. 102724, 2020.
- [18] Linda Wang, Alexander Wong, and Zhong Qui Lin, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *arXiv preprint arXiv:2003.09871*, 2020.
- [19] Joseph Paul Cohen, Paul Morrison, and Lan Dao, "Covid-19 image data collection," *arXiv preprint arXiv:2003.11597*, 2020.
- [20] Jingya Liu, Liangliang Cao, Oguz Akin, and Yingli Tian, "Accurate and robust pulmonary nodule detection by 3d feature pyramid network with self-supervised feature learning," *arXiv preprint arXiv:1907.11704*, 2019.
- [21] Sunghwan Yoo, Isha Gujrathi, Masoom A Haider, and Farzad Khalvati, "prostate cancer detection using deep convolutional neural networks," *Scientific Reports*, vol. 9, 2019.
- [22] To Dat, Dinh Thi Lan, Thi Thu Hang Nguyen, Thi Thuy Nga Nguyen, Hoang-Phuong Nguyen, Le Phuong, and Tien Zung Nguyen, "Ensembled skin cancer classification (isic 2019 challenge submission)," 2019.
- [23] Biraja Ghoshal and Allan Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.
- [24] Daniel Kermany, Kang Zhang, and Michael Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, 2018.
- [25] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [27] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.
- [28] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [30] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [31] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [32] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.
- [33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception

- architecture for computer vision,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [36] Bolei Zhou, Yiyu Sun, David Bau, and Antonio Torralba, “Revisiting the importance of individual units in cnns via ablation,” *arXiv preprint arXiv:1806.02891*, 2018.
- [37] Igor Kononenko et al., “An efficient explanation of individual classifications using game theory,” *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 1–18, 2010.
- [38] Avanti Shrikumar, Jocelin Su, and Anshul Kundaje, “Computationally efficient measures of internal neuron importance,” *arXiv preprint arXiv:1807.09946*, 2018.
- [39] Klas Leino, Shayak Sen, Anupam Datta, Matt Fredrikson, and Linyi Li, “Influence-directed explanations for deep convolutional networks,” in *2018 IEEE International Test Conference (ITC)*. IEEE, 2018, pp. 1–8.
- [40] Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [41] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *arXiv preprint arXiv:1312.6034*, 2013.
- [42] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne, “Investigating the influence of noise and distractors on the interpretation of neural networks,” *arXiv preprint arXiv:1611.07270*, 2016.
- [43] Aravindh Mahendran and Andrea Vedaldi, “Salient deconvolutional networks,” in *European Conference on Computer Vision*. Springer, 2016, pp. 120–135.
- [44] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [45] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.
- [46] Avanti Shrikumar, Peyton Greenside, and Anshul 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.
- [47] Utku Ozubulak, “Pytorch cnn visualizations,” <https://github.com/utkuozbulak/pytorch-cnn-visualizations>, 2019.
- [48] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [49] Grigorios Tsoumakas and Ioannis Katakis, “Multi-label classification: An overview,” *International Journal of Data Warehousing and Mining (IJDDWM)*, vol. 3, no. 3, pp. 1–13, 2007.
- [50] Francisco Charte, Antonio J Rivera, María J del Jesus, and Francisco Herrera, “Addressing imbalance in multilabel classification: Measures and random resampling algorithms,” *Neurocomputing*, vol. 163, pp. 3–16, 2015.
- [51] David W. Cleary Denise E. Morris and Stuart C. Clarke, “Secondary bacterial infections associated with influenza pandemics,” *arXiv preprint arXiv:1907.11704*, 2017.
- [52] Kyle Y. Carver Shigeo Hanada, Mina Pirzadeh and Jane C. Deng, “Respiratory viral infection-induced microbiome alterations and secondary bacterial pneumonia,” *Front. Immunol.*, 2018.
- [53] Andreas Krug and Sebastian Stober, “Visualizing deep neural networks for speech recognition with learned topographic filter maps,” *arXiv preprint arXiv:1912.04067*, 2019.
- [54] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.
- [55] Scott M Lundberg and Su-In Lee, “A unified approach to interpreting model predictions,” in *Advances in neural information processing systems*, 2017, pp. 4765–4774.



COVID-19

Fig. 8. Comparison of various interpretability techniques with respect to models for COVID-19 predictions



Pneumonia

Fig. 9. Comparison of various interpretability techniques with respect to models for pneumonia predictions



Fig. 10. Comparison of various interpretability techniques with respect to models for viral pneumonia predictions

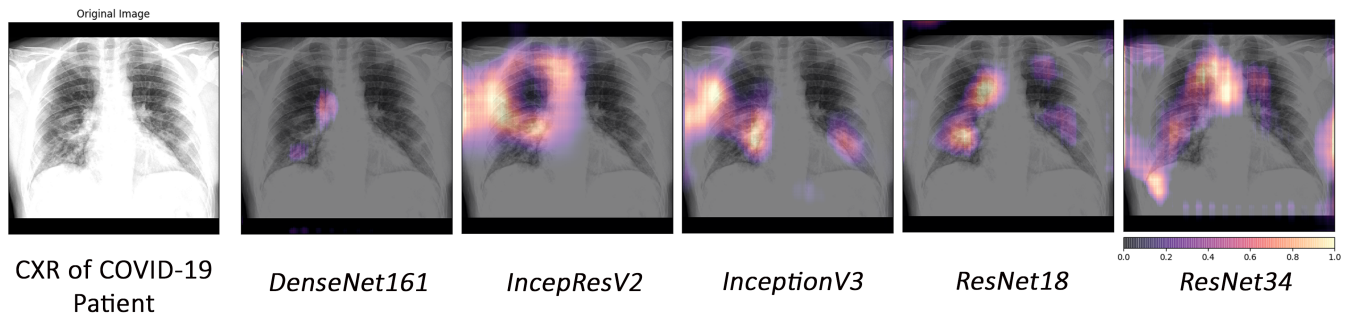


Fig. 11. A case-study of DenseNet161 failure using occlusion

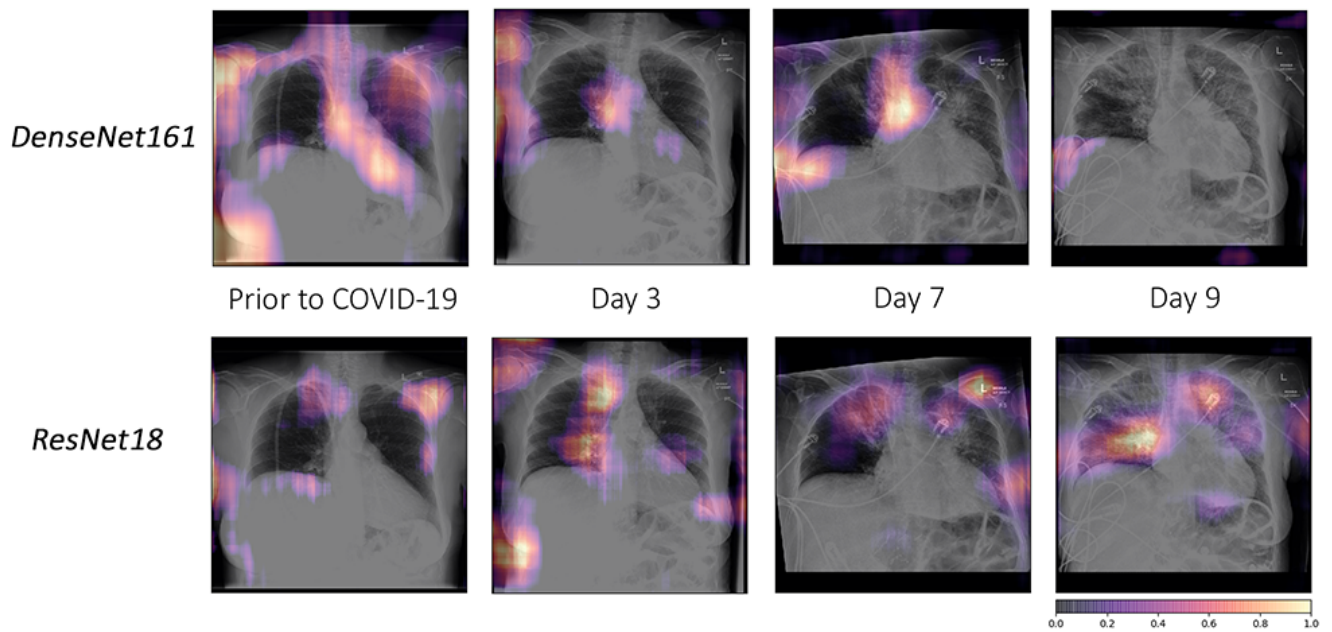


Fig. 12. Comparison using occlusion between DenseNet161 (top) and ResNet18 (bottom) for a specific COVID-19 follow-up case

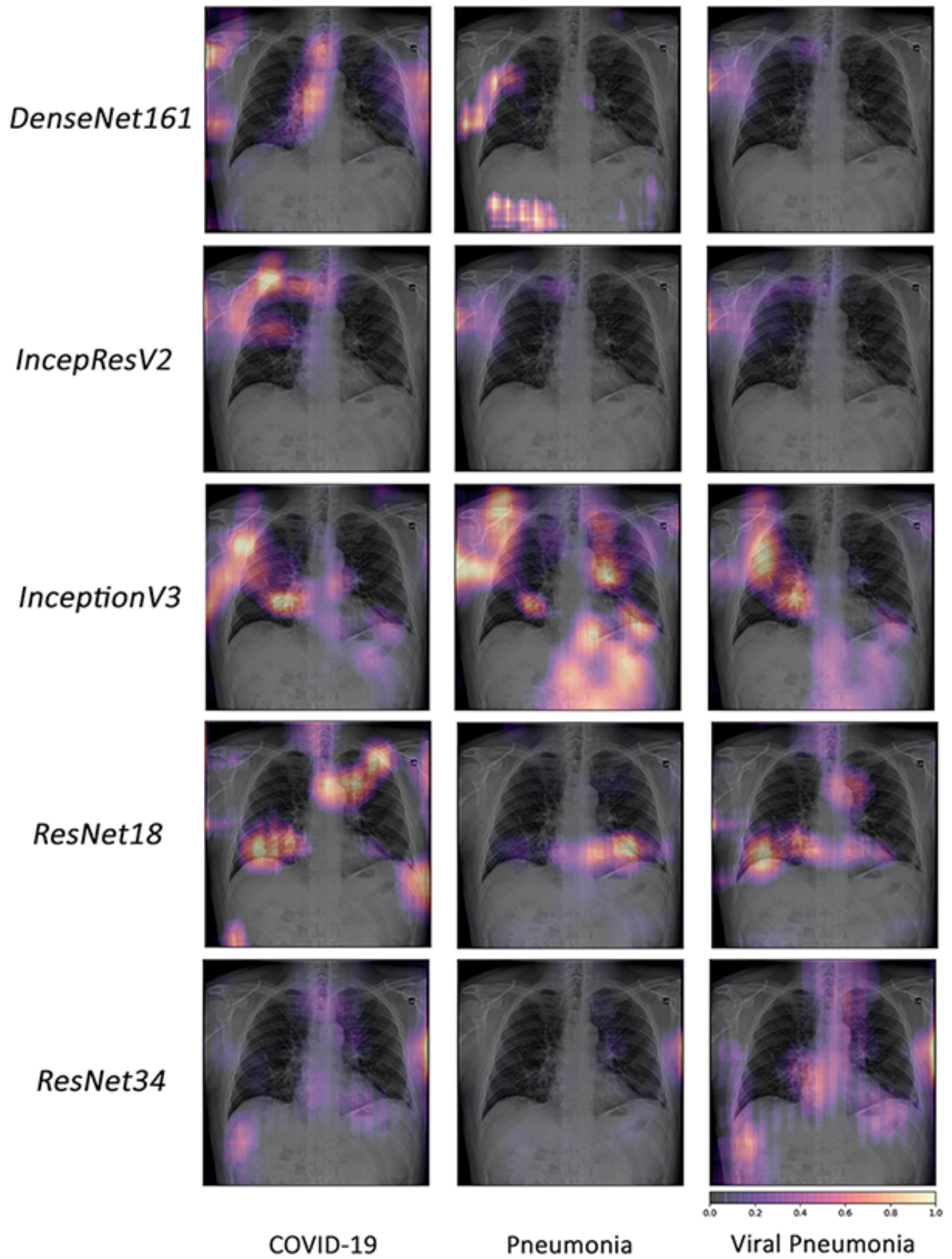


Fig. 13. Example of occlusion for lung pathologies: COVID-19, pneumonia and viral pneumonia