Journal of
Imaging

IMPACT FACTOR 3.2

Indexed in: PubMed

CITESCORE 4.4

Article

Exploration of Interpretability Techniques for Deep COVID-19 Classification Using Chest X-ray Images

Soumick Chatterjee, Fatima Saad, Chompunuch Sarasaen, Suhita Ghosh, Valerie Krug, Rupali Khatun, Rahul Mishra, Nirja Desai, Petia Radeva, Georg Rose et al.

MDPI

https://doi.org/10.3390/jimaging10020045

Journal of
Imaging

MDPI

Article

Exploration of Interpretability Techniques for Deep COVID-19 Classification Using Chest X-ray Images

Soumick Chatterjee [1,2,3,*,†], Fatima Saad [4,5,†], Chompunuch Sarasaen [4,5,6,†], Suhita Ghosh [2,7,†], Valerie Krug [2,7], Rupali Khatun [8,9], Rahul Mishra [10], Nirja Desai [11], Petia Radeva [8,12], Georg Rose [4,5,13], Sebastian Stober [2,7], Oliver Speck [5,6,13,14] and Andreas Nürnberger [1,2,13]

[1] Data and Knowledge Engineering Group, Otto von Guericke University, 39106 Magdeburg, Germany; andreas.nuernberger@ovgu.de
[2] Faculty of Computer Science, Otto von Guericke University, 39106 Magdeburg, Germany
[3] Genomics Research Centre, Human Technopole, 20157 Milan, Italy
[4] Institute for Medical Engineering, Otto von Guericke University, 39106 Magdeburg, Germany
[5] Research Campus STIMULATE, Otto von Guericke University, 39106 Magdeburg, Germany
[6] Biomedical Magnetic Resonance, Otto von Guericke University, 39106 Magdeburg, Germany
[7] Artificial Intelligence Lab, Otto von Guericke University, 39106 Magdeburg, Germany
[8] Department of Mathematics and Computer Science, University of Barcelona, 08028 Barcelona, Spain
[9] Translational Radiobiology, Department of Radiation Oncology, Universitätsklinikum Erlangen, 91054 Erlangen, Germany
[10] Apollo Hospitals, Bilaspur 495006, India
[11] HCG Cancer Centre, Vadodara 390012, India
[12] Computer Vision Centre, 08193 Cerdanyola, Spain
[13] Centre for Behavioural Brain Sciences, 39106 Magdeburg, Germany
[14] German Centre for Neurodegenerative Diseases, 39106 Magdeburg, Germany
[*] Correspondence: contact@soumick.com
[†] These authors contributed equally to this work.

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 diagnosing 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 medical personnel in the diagnosis process. Thus, in this article, five different deep learning models (ResNet18, ResNet34, InceptionV3, InceptionResNetV2, and DenseNet161) and their ensemble, using majority voting, have been used to classify COVID-19, pneumoniæ and healthy subjects using chest X-ray images. Multilabel classification was performed to predict multiple pathologies for each patient, if present. Firstly, the interpretability of each of the networks was thoroughly studied using local interpretability methods—occlusion, saliency, input X gradient, guided backpropagation, integrated gradients, and DeepLIFT—and using a global technique—neuron activation profiles. The mean micro F1 score of the models for COVID-19 classifications ranged from 0.66 to 0.875, and was 0.89 for the ensemble of the network models. The qualitative results showed that the ResNets were the most interpretable models. This research demonstrates the importance of using interpretability methods to compare different models before making a decision regarding the best performing model.

Keywords: COVID-19; pneumonia; chest X-ray; multilabel image classification; deep learning; model ensemble; interpretability analysis

check for updates

Citation: Chatterjee, S.; Saad, F.; Sarasaen, C.; Ghosh, S.; Krug, V.; Khatun, R.; Mishra, R.; Desai, N.; Radeva, P.; Rose, G. Exploration of Interpretability Techniques for Deep COVID-19 Classification Using Chest X-ray Images. *J. Imaging* **2024**, *10*, 45. https://doi.org/10.3390/jimaging10020045

Academic Editor: Antonio Fernández-Caballero

Received: 8 January 2024
Revised: 24 January 2024
Accepted: 5 February 2024
Published: 8 February 2024

1. Introduction

In 2020, the world witnessed a serious new global health crisis, the outbreak of the infectious COVID-19 disease, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1,2]. Due to its long incubation period and its highly contagious nature, it is important to identify infected cases early and isolate them from the healthy population. To date, viral nucleic acid detection using reverse transcription polymerase chain reaction (RT-PCR) has been regarded as the gold standard diagnostic method [3]. However, RT-PCR tests have been reported to suffer from a high rate of false negatives owing to laboratory and sample collection errors [4,5].

However, medical imaging emerges as a great alternative candidate for screening COVID-19 cases and discriminating them from other conditions, as the majority of infected patients exhibit abnormalities on medical chest imaging [6–8]. In this context, chest radiography (CXR) and computed tomography (CT) are widely utilised in front-line hospitals for diagnosis [9–11]. In certain instances, chest CT images have been demonstrated to exhibit higher sensitivity than RT-PCR and have detected COVID-19 infections in patients with negative RT-PCR results [4,11–13]. Nevertheless, there are numerous advantages to encourage the use of CXR imaging in clinical practice, such as faster diagnosis, infection control, and less harmfulness than CT [14,15]. Moreover, X-ray machines are far more readily available than CT scanners, especially in developing countries. In addition, with the help of portable X-ray machines, imaging can be performed in 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 [16]. Despite its limitations, CXR is more widely available than CT across the globe and is widely utilised for COVID-19 screening [16].

Airspace opacities or ground-glass opacities (GGO) are commonly reported radiological appearances with COVID-19 [17,18]. Predominant distributions in the bilateral, peripheral, and lower zones are primarily observed (90%) [19]. However, these manifestations are very similar to various viral pneumoniæ and other inflammatory and infectious lung diseases. Therefore, it is difficult for radiologists to discriminate COVID-19 from other types of pneumoniæ [20]. Expert radiologists are needed to achieve high diagnostic performance, and the duration of the diagnostic is relatively long.

Artificial intelligence (AI) can play one of the potential roles in strengthening the power of imaging tools to provide an accurate diagnosis. Many AI applications have focused on infection quantification and identification to assist radiologists in decision making. The classification of COVID-19 and other types of pneumonia has been investigated using deep learning techniques [6,21]. However, due to the "black box" nature, the rationale behind such techniques is often unknown; hence, these techniques are considered to have low reliability to be integrated within the clinical workflow. Interpretability techniques, which show the focus area of such deep learning methods, are potentially needed to build the confidence of medical practitioners in such methods. Techniques have been proposed that also involve interpretability to understand the reasoning performed by the model [22]. However, comparative studies of different models based on accuracy and interpretability, and then verification of the interabilities by doctors. have not been performed. 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 involving doctors 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 multilabel classification was performed—a task not commonly performed in similar studies. The motivation behind considering deep learning and not interpretable non-deep learning techniques is owing to the fact that, in recent times, deep learning techniques have been observed to outperform others for various radiological applications [23–25].

The remainder of the paper is organised as follows: in Section 1, several related works are presented and discussed, followed by Section 2, which details the various network models and interpretability techniques used here, and the approach to dataset creation is delineated. Section 3 presents the classification results and the interpretability analysis. The results are then analysed in Section 4, and, finally, Section 5 concludes the work and provides directions for further research.

Related Works

The use of artificial intelligence (AI) in healthcare has been developed to support humans in decision making [26–29]. AI-based knowledge has been combined with medical imaging to enhance the accuracy of diagnoses of various diseases, such as respiratory infectious diseases [30] and pulmonary tuberculosis [31], including pandemic diseases such as H1N1 influenza [32].

The spread of COVID-19 has attracted many researchers to concentrate their efforts toward developing AI-based disease detection techniques for various medical imaging modalities. The assistance of deep learning has shown an improvement in binary diagnosis (presence or absence of COVID-19) from CXR images [33] and a reduction in the workload of front-line radiologists [34]. Many efforts have been made to perform multiclass classification (COVID-19, other types of pneumonia, or healthy) to assist radiologists in decision making. Narin et al. [7] used ResNet50, InceptionV3, and InceptionResNetV2 models to classify patients with COVID-19 using CXR images. They demonstrated that the pre-trained ResNet50 model yields the highest accuracy (98%). However, accuracy is often deemed a misleading metric in the case of imbalanced datasets. Furthermore, they only discriminated between healthy subjects and COVID-19, but did not include the other types of pneumonia. Wang et al. [35] designed COVID-Net using CXR images for the classification of patients with bacterial pneumonia, viral pneumonia, COVID-19, and also healthy subjects with a sensitivity of detection of 91% COVID-19. Zhang et al. [6] used a ResNet-based model to classify COVID-19 and non-COVID-19 patients. They achieved a sensitivity of 96% and a specificity of 70.7%. Ghoshal et al. [36] presented a drop-weight-based Bayesian convolutional neural network (BCNN) for CXR-based COVID-19 diagnosis. They found a drastic correlation between the accuracy of the prediction and the uncertainty of the model. Awareness of diagnosis decision uncertainty could endorse deep learning-based applications to be used more and more in clinical routines. Singh et al. [37] proposed the Gen-ProtoPNet architecture that provides interpretable classifications of COVID-19 in CXR [37] and CT scans [38], resulting in F1 scores as high as 98%. Furthermore, Shorten et al. [39] provided a comprehensive survey of different applications of deep learning for COVID-19. On the other hand, De Falco et al. [40] proposed an interpretable, completely transparent evolutionary rule-based approach, but only managed to achieve an accuracy of around 80%. This demonstrates the possible trade-off between transparency and model performance. Deep learning methods that are interpretable, or that are interpreted using post hoc methods, can mitigate this trade-off. Although the application of deep learning methods for COVID-19 lesion detection is not an unexplored topic, including interpretability, systematic comparisons of different models in terms of interpretability and verification of the interpretability results by medical professionals are still missing. These are the aspects this paper seeks to address, while presenting the importance of evaluating or comparing models with respect to interpretability along with classification accuracy. It is noteworthy that these problems and the message of this paper are not limited to COVID-19 classification, but are also applicable to classification problems in general, especially in high-risk domains like medical imaging.

Although AI-based assistance has been present in the field of radiology for a long time, the decision-making mechanisms within these “black-box” methods remain questionable. Recently, research on interpretability has gained more focus. Different interpretability techniques, such as occlusion [41], saliency [42], guided backpropagation [43], integrated gradients [44], etc., have been introduced, demonstrating the potential to open these black boxes.

2. Materials and Methods

2.1. Network Models

During the course of this research, various network architectures were explored and experimented with, including several variants of VGG [45], ResNet [46], ResNeXt [47], WideResNet [48], Inception [49], and DenseNet [50]. Prior to training on the dataset of this research work, all the networks were initialised with weights pre-trained on ImageNet.

After observing the results, five network architectures were shortlisted for further analysis and also used to create an ensemble using the majority voting strategy for better prediction performance. The models were selected based on different criteria, such as performance, complexity of the model, etc. The selected models are discussed in this section, and Table 1 shows the complexity of the models.

Table 1. Number of trainable parameters in each model.

Model	No of Parameters	GFLOPs	MACs ($\times 10^9$)	GPU Memory (Forward + Backward) in GB
ResNet18	11,183,694	18.95	9.53	0.15
ResNet34	21,291,854	38.28	19.22	0.22
InceptionV3	24,382,716	35.04	17.63	0.44
DenseNet161	26,502,926	80.73	40.98	1.31
InceptionResNetV2	54,327,982	81.07	40.70	0.72

ResNet:

At the nascent stage of deep learning, the deeper networks faced the problem of vanishing gradients/exploding gradients [51,52], which hampered convergence. The deeper network faced another obstacle called degradation, where the accuracy starts to saturate and degrade rapidly after a certain depth of the network. To overcome these problems, He et al. [46] 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 hyperparameter or learnable parameter but just adding the output from a preceding layer to a subsequent layer. It unleashed the possibility of training deeper models whilst avoiding these aforementioned issues.

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

InceptionNet:

An image can have thousands of salient features. In different images, the focused features can be in any different part of the image, which makes determining the appropriate kernel size for a convolution network a very difficult task. A large kernel has a greater focus on globally distributed information, while a smaller kernel focuses on local information. To overcome this problem, Szegedy et al. [49] 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 made the network "wider" rather than "deeper". In order to enhance computational cost-effectiveness, the authors restricted the number of input channels by adding an extra 1 × 1 convolution before the 3 × 3 and 5 × 5 convolutions. Adding 1 × 1 convolutions is much cheaper than adding 5 × 5 convolutions. The authors introduced two auxiliary classifiers to avoid the problem of a 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.

Excessive reduction in dimensions can cause a loss of information, also known as a "representational bottleneck". To overcome this problem and scale the network in ways that utilise the added computation as efficiently as possible, the authors of InceptionNet introduced a new idea in another publication by Szegedy et al. [53], factorising convolutions and aggressive regularisation. The authors factored each 5 × 5 convolution into two 3 × 3 convolution operations to improve computational speed. Furthermore, they factorised the convolutions of the filter size nxn into a combination of the 1 × n and n × 1 convolutions. This network is known as InceptionV2.

Szegedy et al. [53] also proposed InceptionV3, which extends InceptionV2 further by factorising 7 × 7 convolutions, label smoothing, and by adding BatchNorm in the auxiliary classifiers. Label smoothing is a type of regularising component added to the loss formula that prevents the network from becoming too confident about a class.

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

InceptionResNetV2:

The different variants of InceptionNet and ResNet have shown very good performance with relatively low computational costs. With the hypothesis that residual connections would cause Inception network training to accelerate significantly, the authors of the original InceptionNet proposed InceptionResNet [54]. 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 (1 × 1 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.

DenseNet:

Huang et al. [50] came up with a very simple architecture to ensure maximum information flow between layers of 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, such as ResNet, DenseNets do not sum the output feature maps of the layer with the incoming feature maps but concatenate them.

In the preliminary trials of this study, DenseNet161 came out as a winner in terms of performance. Therefore, DenseNet161 was included in this research.

2.2. Interpretability Techniques

Interpretability techniques can aid in understanding the reasoning of a network for its predictions. In general, the results of interpretability can be visualised using heatmaps, where higher values indicate a heightened focus. However, this may vary among different interpretability techniques. Typically, the heatmaps are overlaid on top of an input image to understand at which parts of the image the network is focused to generate the predictions. The techniques that use a single image at a time for analysis are known as local interpretability techniques. On the other hand, a global interpretability technique often pertains to comprehending how the model works—an aggregated behaviour of the model based on the distribution of the data [55,56]. There are several techniques already in existence. Some of the methods, such as occlusion, saliency, input X gradient, integrated gradients, guided backpropagation, DeepLIFT, and neuron activation profiles, which have been explored in this research, are explained briefly in this section.

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 obtain this answer, Zeiler et al. [41] performed an occlusion technique by systematically blocking different parts 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 subsequently feeds into the trained network to obtain probability scores for a given class for each mask position.

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 deep learning models. The saliency method of Simonyan et al. [42] returns the gradients of a model for its respective inputs. Positive values present in the gradients show how a small change in the input image changes the prediction.

Input X Gradient:

Input X gradient is an extension of the saliency approach. Similarly to the saliency method of Simonyan et al. [42], this method of Kindermans et al. [57] also takes the gradients of the output with respect to the input, but additionally multiplies the gradients by the input feature values.

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 positive error signals are backpropagated, and the negative signals are set to zero while backpropagating through a ReLU unit [58].

Integrated Gradients:

Sundararajan et al. [44] proposed a model interpretability technique, which assigns an importance score to each of the features of the input by approximating the integral of the gradients of the output for that input, along the path from the given references for the input.

DeepLIFT:

Deep Learning Important FeaTures or DeepLIFT, proposed by Shrikumar et al. [59], is a method to pixel-wise decompose the output prediction of a neural network on a specific input. This involves 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 that 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 where the gradient is set to zero.

Neuron Activation Profiles:

The aforementioned interpretability techniques are local methods that help to understand single predictions of a neural network. To investigate model behaviour more generally, a global interpretability technique called neuron activation profiles (NAPs) is employed [60,61]. NAPs describe and contrast the activity of the neural network of sets of related inputs, for example, of different classes, using an averaging approach. Initially, the activation values in the layers of interest are obtained by computing a forward pass for every test image. Then, the average feature maps over each respective group are computed to characterise the group-specific activity. In addition to characterising the network activations for a group, further emphasis is given to the differences between the groups. To this end, the average over all groups is subtracted from each group's average. These normalised averaged activation values can be interpreted as the activation difference from the global average. Positive values indicate a characteristically high neuron activation compared with the entire dataset, and negative values indicate a comparably low neuron activation. NAP values are particularly useful to identify which activations differ between groups of interest and correspondingly indicate the model's ability to distinguish between the classes according to the activations. When working with image data, visually interpretable plots of NAPs of feature maps can be created. For data that are not visually interpretable, NAPs can be further used for similarity analyses [61] or for dimensionality reduction-based visualisation [62].

In order to obtain useful averaging results, this method requires data in which the objects are at the same location in the images. This alignment is guaranteed through data preprocessing that resizes and crops the original images.

2.3. Implementation

The models were implemented using PyTorch [63]. An interpretability pipeline for PyTorch-based classification models was developed with the help of Captum [64]. The code of this project is available on GitHub: https://github.com/soumickmj/diagnoPP. The pipeline was later made part of the TorchEsegeta [65].

Training sessions were conducted using Nvidia GeForce 1080 Ti and 2080 Ti GPUs, each with 11GB of memory. The loss was calculated using binary cross-entropy (BCE) with logits, which combines the 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 minimised by optimising the model parameters using the Adam optimiser [66], with a learning rate of 0.001 and a weight decay of 0.0001. A manual seed was used to ensure the reproducibility [67] of the models. Automatic mixed precision was used using Apex [68] to speed up training and decrease GPU memory requirements.

The interpretability methods were applied on the models using Nvidia Tesla V100 GPUs, having 32GB memory each. Some of the interpretability techniques could not be used on certain models owing to insufficient GPU memory caused by the complexities of the models.

2.4. Data

2.4.1. Data Collection

The CXR images were collected from two public datasets. The first dataset was the COVID-19 image data collection by Cohen et al. [21,69], comprising 236 images of COVID-19, 12 images of COVID-19 and ARDS, 4 images of ARDS, 1 image of Chlamydophila, 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 Kermany et al. [70,71], which 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 subjects, 250 images of viral pneumonia, and 250 images of bacterial pneumonia were randomly chosen. Figure 1 portrays the final data distribution considered for the work. This CXR image dataset comprises posterior anterior (PA), anterior superior (AP), and anterior superior supine (AP supine) radiographs. Whilst the AP view is not the preferred positioning and has disadvantages such as organ overlap that could interfere with network prediction [72], it is a technique commonly used for COVID-19 patients in a coma.

The hierarchical nature of the pathologies can be observed in this combined dataset (see Figure 2). For example, SARS and COVID-19 are subtypes of viral pneumonia. However, Streptococcus, Klebsiella, Chlamydophila, and Legionella are subtypes of bacterial pneumonia, and Pneumocystis is a subtype of fungal pneumonia. Furthermore, viral, bacterial, and fungal pneumoniæ are different types of pneumonia. Therefore, a patient with COVID-19 inherently has viral pneumonia. ARDS, which stands for acute respiratory distress syndrome, is a serious lung condition with a high mortality rate [73]. It frequently develops alongside pathological conditions like nonpulmonary sepsis, aspiration, or pneumonia [74]. Although the respiratory pathologies of ARDS (associated with or without COVID-19) and COVID-19 are similar, COVID-19 has different features that require different patient management, and a patient suffering from both could require additional care [75–77]. Therefore, the dataset, which comprises cases where a patient has both COVID-19 and ARDS, is suitable for multilabel classification.

2.4.2. Dataset Preparation

The final dataset was randomly divided into a training set, consisting of 60% of unique subjects and the remaining 40% of the subjects being used as a test set. Five-fold cross-validation (CV) was conducted to assess the generalisation capabilities of the models. The performance of the models during the 5-fold CV is reported in Section 3.1. For the interpretability analysis, only the results from the first fold were used, as this yielded the highest micro F1 scores.

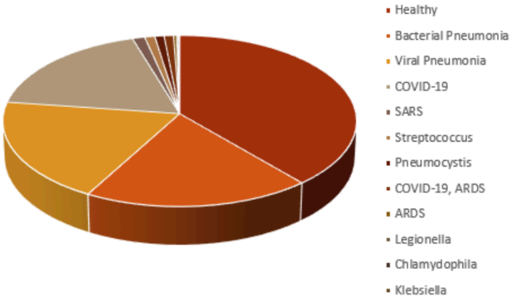


Figure 1. CXR images distribution for each infection type in the dataset.

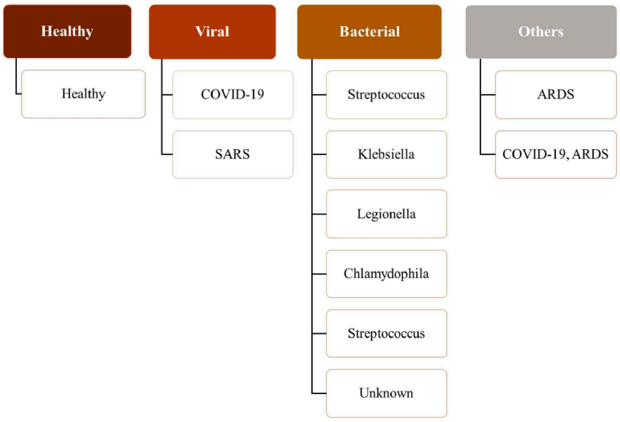


Figure 2. A hierarchy of pathological labels used in this study.

2.4.3. Pre-Processing

The dataset used for the task comprises X-ray images collected at different centres using different protocols and varying in size and intensity. Therefore, all the images were initially pre-processed to have the same size. To make the image size uniform throughout the dataset, each image was interpolated employing bicubic interpolation to have 512 pixels on the longer side. The pixel count on the shorter side was determined, keeping the aspect ratio of the original image. Subsequently, zero-padding was applied to the shorter side to make that side have 512 pixels, resulting in a 512 × 512 image. Image resizing was followed by percentile cropping, where the image intensity was cropped to the first and 95th percentile, and then the intensity normalisation was performed to the range [0,1]. The percentile cropping normalisation minimises the effect of intensity variation due to non-biological factors.

2.4.4. Classification Setup

In this multilabel classification setup, the model was trained to identify the disease and also its supertypes. Therefore, 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 with multiple pathologies, as in this dataset where some patients have both COVID-19 and ARDS, ideally, the network should classify it as pneumonia, viral pneumonia, COVID-19, and ARDS. Interpretability analysis was conducted for each label of each image in the test set.

2.5. Evaluation Metrics

In a multiclass setting, classifiers are generally evaluated with respect to precision, recall, and F1 metrics. In a multilabel classification setting, these metrics are computed in two manners: macro and micro averaging [78].

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

As shown in Equation (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 $\in$ {precision, recall, F1}.

This manner of computation of metrics helps to treat each pathology equally, and the metric values are significantly influenced by the rarer labels.

$$Micro = Metric(\sum_{i=1}^{p} TP_i, \sum_{i=1}^{p} FP_i, \sum_{i=1}^{p} TN_i, \sum_{i=1}^{p} FN_i). \tag{2}$$

In micro-based metrics, TP, TN, FP, and FN of each class/pathology are added individually and then averaged, as shown in Equation (2). Therefore, the micro-based metrics portray the aggregated contribution of all classes/pathologies. Therefore, the influence of the predictions from the minority classes becomes diluted among the contributions from the majority classes. This makes the micro-based metrics a suitable measure for estimating the overall performance of the classifier, particularly in scenarios involving imbalanced datasets. Given the significant imbalance in the utilised dataset, micro-based metrics have been considered for classifier evaluation [79].

3. Results

3.1. Model Outcome

3.1.1. Overall Comparisons of the Classifiers

Figure 3a shows that the overall performance of the classifiers over pathologies was similar. Among the non-ensemble models, DenseNet161 performed the best in all metrics. Although InceptionResNetV2 was the most complex model among all, it yielded the poorest recall, which implies that the ability of the model to find pathology-affected cases was poor compared with less complex models. ResNet18 was the least complex model among the non-ensemble classifiers, ranking second to DenseNet161 with respect to micro F1. The ensemble produced the best results and the minimum variance in the 5-fold cross-validation, as presented in Table 2.

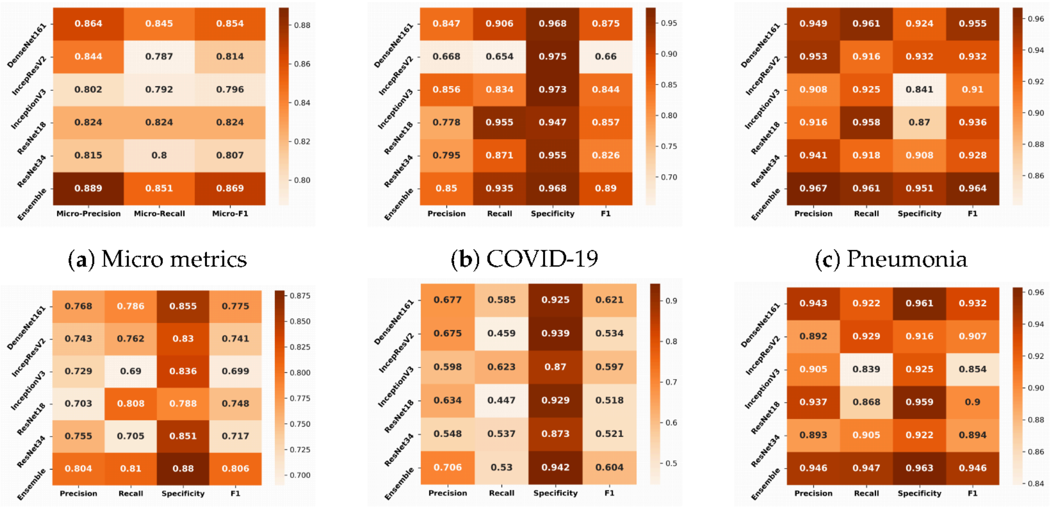
Table 2. Performance of all the classifiers with respect to micro-based metrics over 5-folds.

Model	Precision	Recall	F1
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

Another interesting observation that could be made is regarding inactive feature maps (dead neurons). DenseNet161 had the highest percentage of such feature maps—as high as 99.22% for the middle layer. Although InceptionResNetv2 was the most complex, it had fewer inactive feature maps than DeseNet161. ResNets, the least complex models in this study, had the lowest percentage of inactive feature maps (48.44% and 60.16% for the middle layers of ResNet18 and ResNet34, respectively).

3.1.2. Comparisons of the Classifiers for Different Pathologies

The authors also compared the classifiers' performance at the pathology level. The average metric values across five cross-validation folds are depicted in Figure 3b–f for COVID-19, pneumonia, viral pneumonia, bacterial pneumonia, and healthy subjects, respectively. When comparing the models using the average F1, it was observed that the performance of most models for COVID-19, pneumonia, and healthy was good, except for the performance of InceptionResNetV2 for COVID-19 cases. Among all models, the results of DenseNet161 were the most promising for all diseases. For the COVID-19 classification, DenseNet161 performed the best, and ResNet18 was in second position. DenseNet161 performed the best for pneumonia. InceptionResNetV2 provided the highest performance for the classification of viral pneumonia. Lastly, InceptionV3 gave the highest scores for bacterial pneumonia.



(**a**) Micro metrics (**b**) COVID-19 (**c**) Pneumonia

(**d**) Viral Pneumonia (**e**) Bacterial Pneumonia (**f**) Healthy subjects

Figure 3. Comparison of the classifiers based on micro metrics (**a**) and their performance for the different classes (**b**–**f**).

3.2. Interpretability of Models

In Section 3.2.1, different interpretability techniques are explored for different classifiers with respect to the different diseases. Section 3.2.2 talks about how the different models performed for specific pathologies.

All the given interpretability analyses (except using the global method NAP) were performed for that specific input CXR image that was shown as the underlay. In the interpretability analysis using NAP, all images from the test set were used, as this method performs a global analysis.

3.2.1. Pathology-Based Comparisons of Local Interpretability Techniques for Models

To visualise the results for a specific case, the models were interpreted using local methods: occlusion, saliency, input X gradient, guided backpropagation, and integrated gradients, and are shown in Figures 4–6. Apart from occlusion, the other interpretability techniques failed to run for DenseNet161 due to GPU memory limitations. in DeepLIFT, ResNets faced an additional challenge due to the ReLU operations used "in place" in those models. Models have to be updated to run DeepLIFT on them.

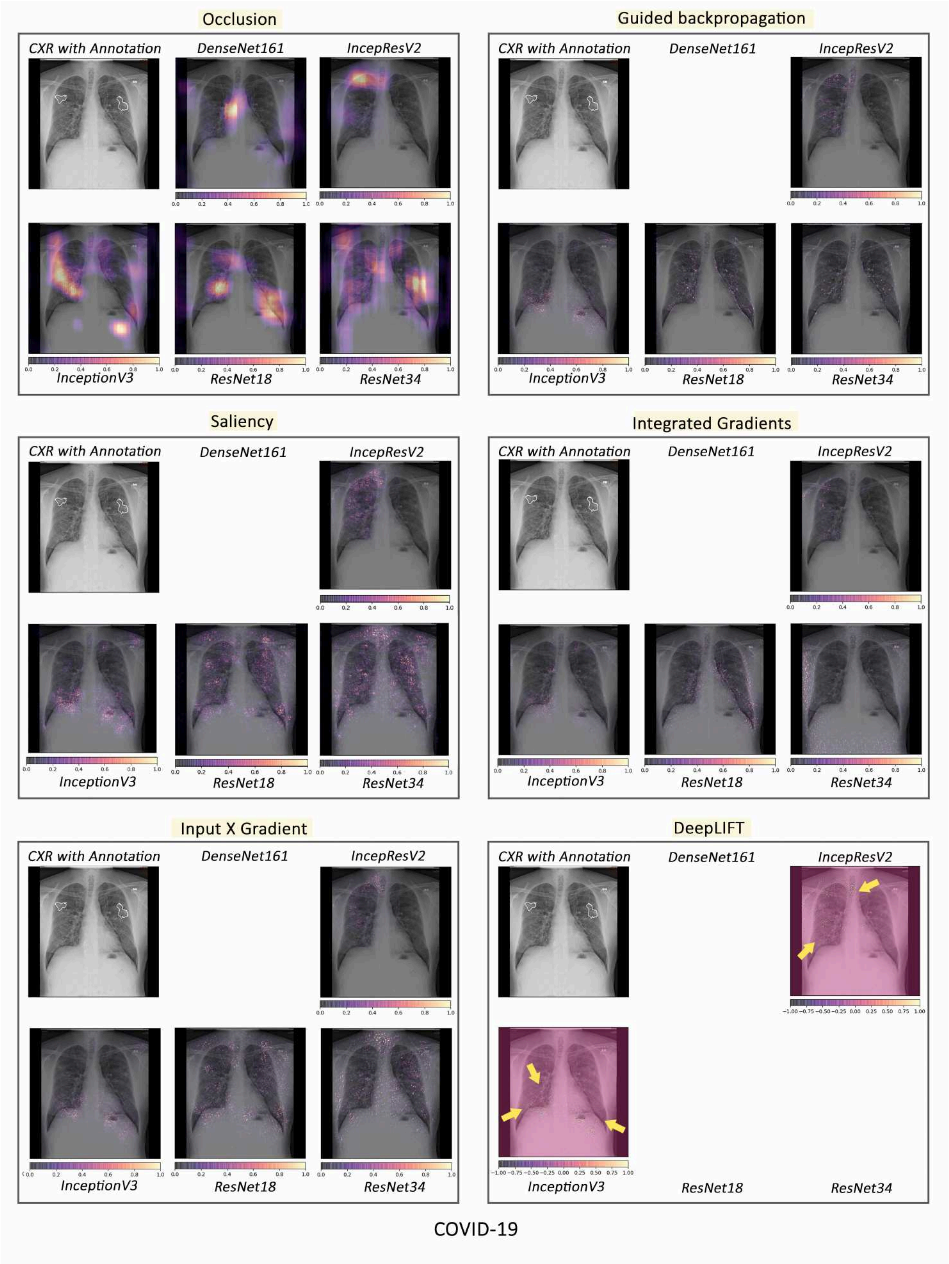


Figure 4. Comparison of various interpretability techniques with respect to models for COVID-19 predictions against the manual annotation of the affected areas by medical experts.

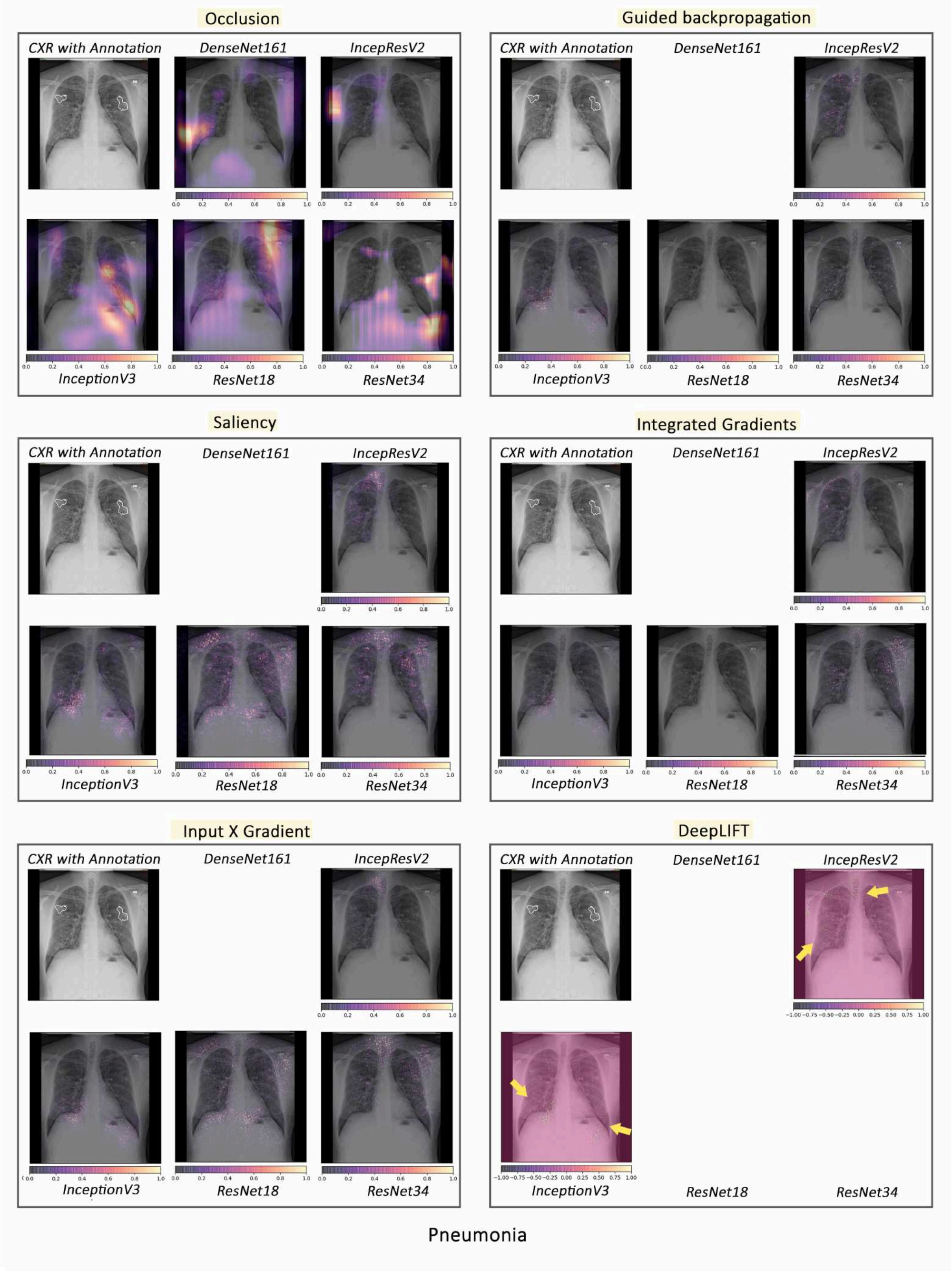


Figure 5. Comparison of various interpretability techniques with respect to models for pneumonia predictions against the manual annotation of the affected areas by medical experts.

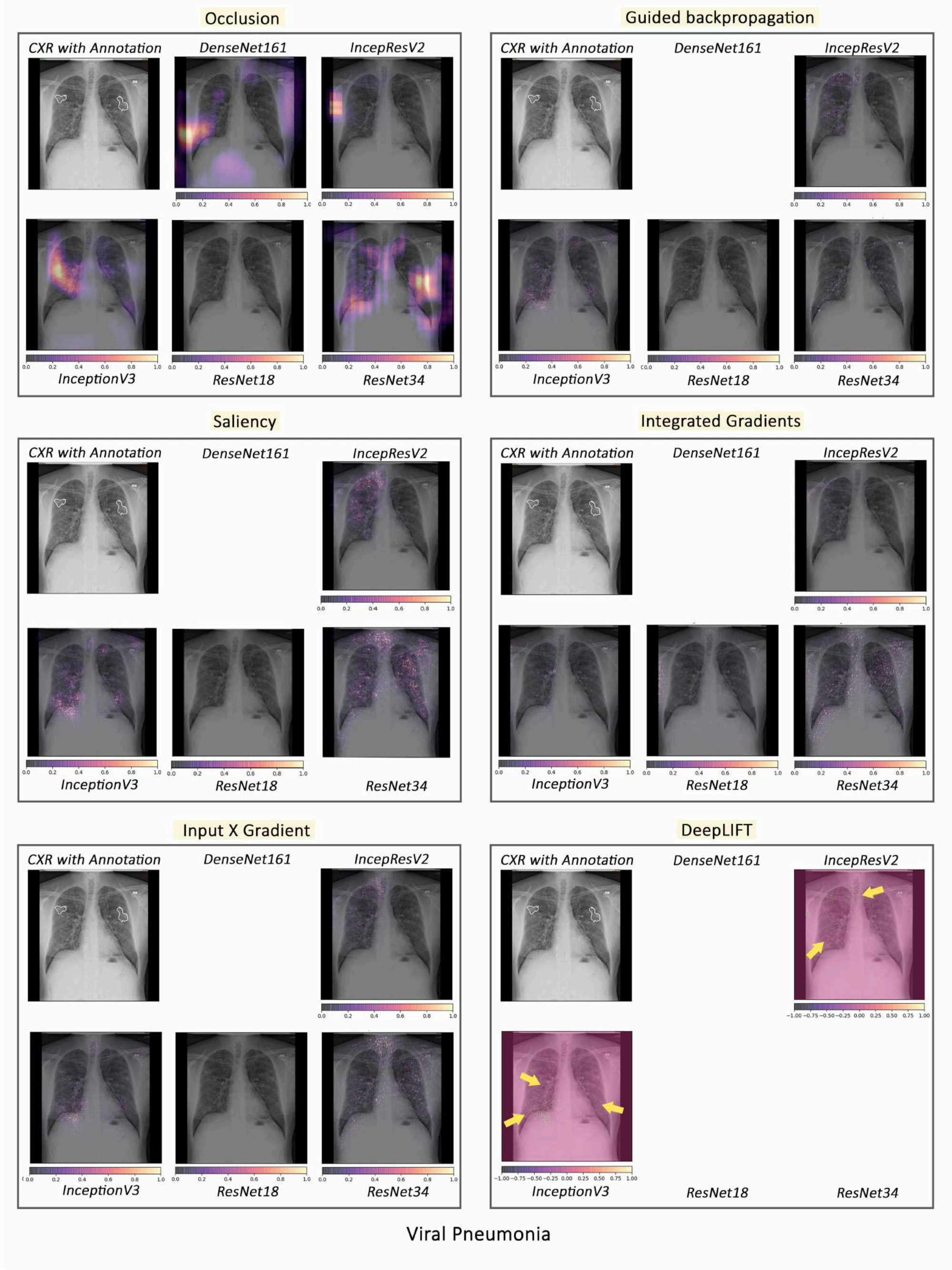


Figure 6. Comparison of various interpretability techniques with respect to models for viral pneumonia predictions against the manual annotation of the affected areas by medical experts.

According to the clinical findings of the COVID-19 image data provided by Cohen et al. [21], multiple abnormalities of the lungs were located in the upper and lower pulmonary field, as well as the upper left part of the lung. The models classified this case as COVID-19, pneumonia, and viral pneumonia responding to the pathology of lung infection. It can be seen that the focus area of the models for COVID-19 differs from the focus area for pneumonia and viral pneumonia. DenseNet161 and InceptionResNetV2 focused primarily on the right lung. InceptionV3, ResNet18, and ResNet34 covered both the right and left parts, not only the lesion but also the irrelevant regions outside the lung.

Local interpretability methods suffered mainly from false positives. In some cases, the occlusion did not detect the affected areas for DenseNet161 and InceptionResNetV2 and falsely marked the normal areas as positive, as shown in Figure 4. Furthermore, for InceptionV3, it detected some positive patches, but falsely detected more areas as positive. Finally, in general, for ResNets, occlusion was most sensitive to positive areas and detected fewer false negatives. Guided backpropagation, saliency, integrated gradients, and DeepLIFT in general falsely detected normal lung areas as positive—they picked up normal bronchovascular markings as positive and did not mark the actual affected areas. The input X gradient detected some positive areas correctly for ResNet18, but falsely marked many normal areas. In general, the representations learnt by the ResNet models captured the most accurate regions, as seen from most interpretability techniques, with fewer false negatives. Among the local interpretability techniques, occlusion provided the best guidance in finding clinically important areas, which were confirmed by medical experts.

3.2.2. Intense Interpretability

The failure case of the best performing model for COVID-19 classification:

Although DenseNet161 performed the best among all models, it gave false negatives for some of the COVID-19 patients, while the rest of the models, including the ensemble, predicted correctly. The occlusion results of the models can be observed in Figure 7. This figure shows that DenseNet161 and InceptionResnetV2 did not focus on any affected areas, but rather on other regions (e.g., normal right hilum). InceptionV3, ResNet18, and ResNet34 mainly focused on affected areas with good sensitivity. InceptionV3, however, had more false positives than ResNets (e.g., outside the right lung).

Another analysis was performed with CXR of a 70-year-old woman who had three days of cough, myalgia, and fever, without any recent overseas travel. A series of chest radiographs were obtained before confirmation of coronavirus infection, and follow-ups were performed at three days, seven days, and nine days, which showed the progression of radiographic changes. In the image prior to COVID-19, both models falsely detected all normal areas as relevant features. In the image of day 3, the doctor could not visually detect any affected area, although this was the image from the third day after testing positive for COVID-19. This may indicate that, when no substantial affected area can be seen in the image visually (i.e., day 3), the model might be picking up some mild markers, which cannot be confirmed visually. In the images of days seven and nine, DesNet161 did not focus correctly on the affected regions and had both false positives and false negatives, while ResNet18 focused on the affected regions more accurately.

ResNet18 can be considered the overall winner, as it yielded high evaluation scores, despite having the least number of network parameters. Furthermore, its interpretability analysis showed the location of the lesion, which allows us to use this network for follow-up or severity estimations, as illustrated in Figure 8.

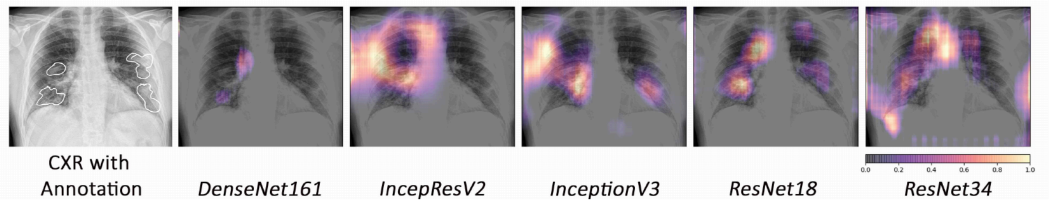


Figure 7. A case-study of DenseNet161 failure using occlusion. The affected areas in the lungs have been annotated by medical experts.

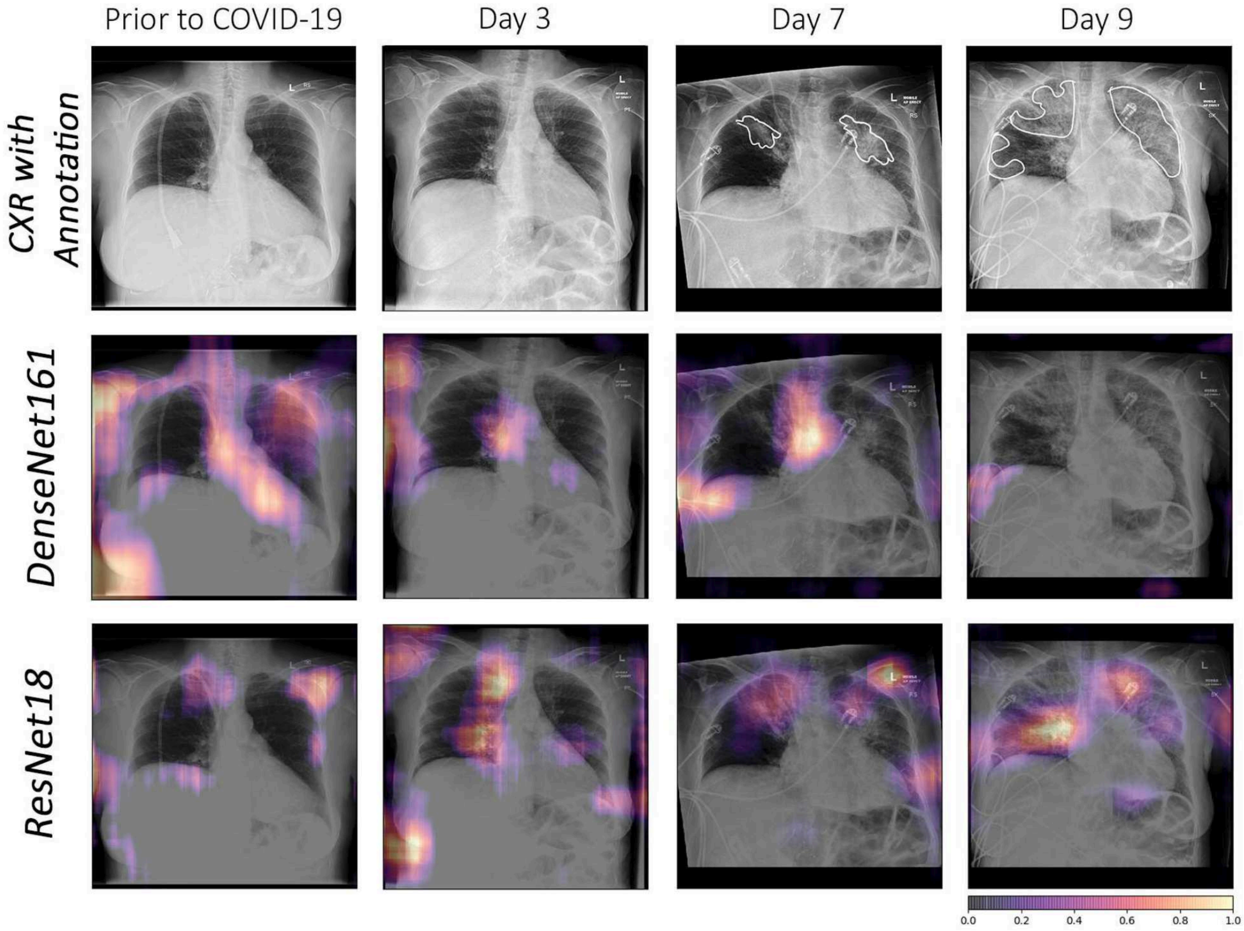


Figure 8. Comparison using occlusion between DenseNet161 and ResNet18 for a specific COVID-19 follow-up case. The affected areas in the lungs have been annotated by medical experts.

Representations in DenseNet161 and ResNet18:

In addition to individual failure cases, the authors further investigated how the COVID-19 and pneumonia pathologies are represented in the neuron activations of DenseNet161 and ResNet18 compared with healthy individuals. This representation analysis was performed using NAPs—a global interpretability technique. In general, in a well-generalised model, larger neuron activation differences are expected between different pathologies and healthy subjects in the lungs than in other image areas. If activity differences are observed in other regions, this indicates that the model exploits biologically irrelevant features to discriminate the classes.

To find potentially exploitable features, the input averages (input layer NAPs) were first investigated in Figure 9 (left). It can be observed that pneumonia images cover a smaller portion of the height dimension than COVID-19 or healthy subject images. This means that there are dark top and bottom regions in the majority of pneumonia images. Based on this observation, the authors hypothesised that a model might exploit this non-biological feature. To investigate this hypothesis, the feature map NAPs of DenseNet161 and ResNet18 in an early and deep layer, respectively, were visualised. The authors particularly investigated layers at representative depths of the networks. For DenseNet161, the ReLU-activated outputs of the first and last dense blocks were chosen. As representative layers of ResNet18, the outputs after the first and last residual connections were selected. For these layers, two exemplary feature map NAPs among those of the highest activity differences between the observed classes are shown in Figure 9. In DenseNet161, one can clearly observe activation differences in both the border regions and the lung. For example, COVID-19 images are easy for the model to distinguish based on the activation difference corresponding to not having dark regions at the bottom and top of the images. In the deeper layer, the activation difference patterns do not resemble any interpretable structure, neither in the lungs nor in the lower and upper regions. This indicates why DenseNet161 has a high performance despite giving false negative COVID-19 results. Instead of detecting COVID-19-specific features, it likely exploits features of the data that are correlated but not related to the pathology. However, it does not appear that DenseNet161 uses dark border regions as the main distinguishing factor. ResNet18, in contrast, is less likely to detect biologically irrelevant features. Although in the early layers there are activation differences in the top and bottom areas of the images, in most deep-layer feature maps, the groups can be most clearly distinguished from each other from neuron activity in the (right) lung regions.

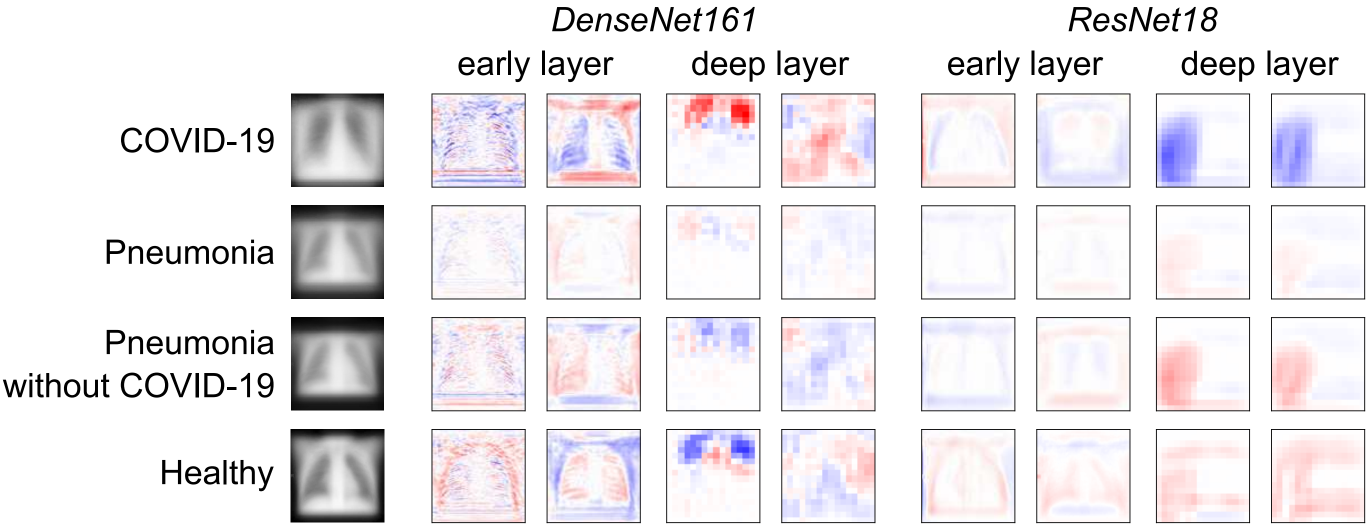


Figure 9. Average input images and feature map NAPs in different models and layers for different pathologies and healthy subjects. Blue indicates lower activation of the respective neuron for this group compared with the other groups, and red indicates higher activity.

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 centralised on the interpretability comparison of the models for these three pathologies. Interpretability techniques reported that different networks focused on different areas for the same CXR image to predict each of the diseases. It was observed that the focus area of DenseNet161 for COVID-19 was explicitly different from that for pneumonia and viral pneumonia. However, InceptionResNetV2 and InceptionV3 emphasised a similar area (different focus areas for each model) for the three pathologies. Furthermore, ResNet18 and ResNet34 targeted the lung region for COVID-19 and viral pneumonia, but differed for pneumonia. Figure 10 exhibits the mentioned findings.

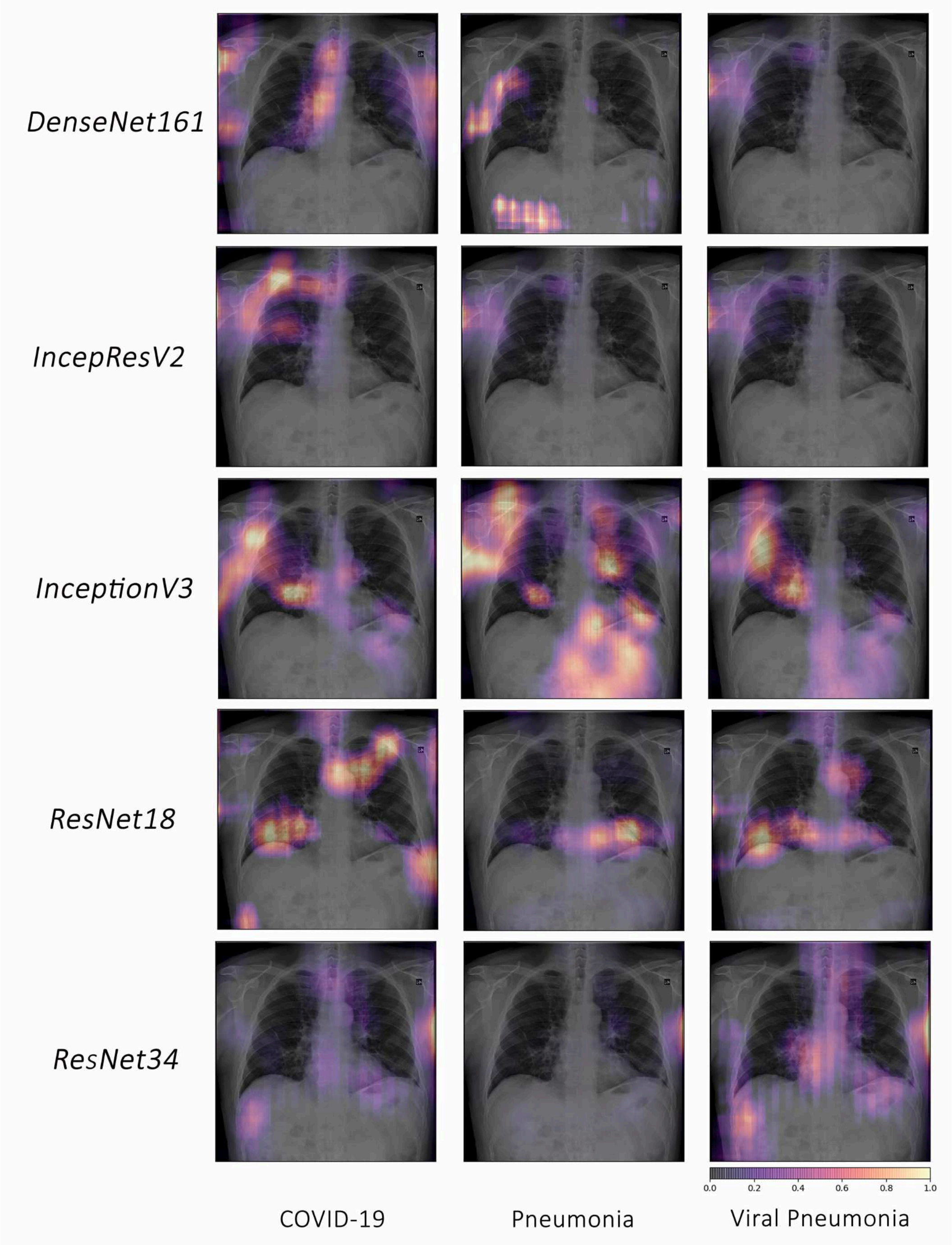


Figure 10. Example of occlusion for lung pathologies: COVID-19, pneumonia, and viral pneumonia.

4. Discussion

The literature review portrays that the diagnosis of COVID-19 is seen as a multiclass classification task rather than a multilabel classification. The datasets used in the previous works vary in terms of the amount of data used for the classification task. In [7], the authors 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 [6,8,35] performed a multiclass classification task on different imbalanced datasets using X-rays, and achieved a maximum mean specificity of 0.989, 0.979, and 0.971, respectively. In this work, InceptionResNetV2 achieved the highest specificity of 0.975, comparable to previous studies. However, in this research, the authors used a different dataset, train–test split, and preprocessing techniques compared with previous works, which makes it unfair to compare the results with previous studies.

It was observed that the less complex models were more interpretable, while having fewer dead neurons than the more complex ones. DeneseNet161, which resulted in the highest F1 score, had the highest number of dead neurons and also had the worst focus areas according to interpretability methods. The model that resulted in the second-best F1 score, ResNet18, was the least complex model in this study—while also having the best focus areas, as dictated by the interpretability methods. This was further confirmed by a global interpretability method, NAPs, which showed that ResNet18 was less likely to detect biologically irrelevant features. It should be noted that, in some cases, the network predicted the findings as a presence of COVID-19, while the doctors did not report any abnormalities.

There were a couple of cases where the network detected both viral and bacterial pneumonia. According to Morris et al. [80] and Shigeo et al. [81], the induction of viral infection could lead to secondary bacterial infection and increase the severity of 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.

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

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

5. Conclusions and Future Works

In this paper, a range of deep learning-based classifiers were compared for the multilabel classification of COVID-19 and similar pathologies in CXR images, and the interpretability of these models was investigated and finally corroborated by medical professionals. In general, most of the models performed well. However, certain models failed at specific tasks. The authors additionally formulated an ensemble employing majority voting, which aided in addressing these models' shortcomings by combining their predictions. Furthermore, the smallest model, ResNet18, was found to compete well with 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 multilabel classification setup was able to predict all of those pathologies correctly. DenseNet161 was the model that performed the best in this setup in terms of classification scores, though it was observed that the focus of the network was often on unrelated biologically irrelevant regions. This can be attributed to the fact that the network discerned some irrelevant patterns in the dataset, which might be due to the high complexity of the model. The highest number of dead neurons was also observed in this model, suggesting that the model may have been overly complex for the given task. After qualitative analysis of the interpretability results, it can be said that the ResNets were the most interpretable models, as the networks focused predominantly on the appropriate regions.

Model explainability methods such as LIME [82], SHAP [83], etc., were not explored during this research but are planned as future work. The 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 nonimage information (like the patient's prior medical history, the result of the RT-PCR test, etc.) could also be integrated into the network models to aid the networks in decision making. Furthermore, instead of supplying the whole image to the models, lung segmentation could serve as a preprocessing step, which might improve the networks' predictions by helping them to focus just on the region of interest, which in this case is the lungs. Training techniques such as 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 multilabel classification problem. Several interpretation techniques were implemented in the interpretability pipeline, but were not investigated in this study and will be explored in the future for this dataset model setup. Finally, in the future, a large-scale study involving more medical professionals should also be performed to evaluate the benefits of interpretability methods in terms of building trust, and also their usefulness in the clinical workflow should also be evaluated in the future.

Author Contributions: S.C., F.S., C.S. and S.G. created the concept and designed the study, under the supervision of G.R., S.S, O.S. and A.N. S.C. and S.G. performed the experiments. V.K. created the neuron activation profiles and analysed their results. C.S. performed the qualitative analysis of the interpretability results and created the visualisations. R.M. and N.D. reviewed the interpretability results and created the annotations. S.C., F.S., C.S., S.G. and R.K. wrote the manuscript. P.R., G.R., S.S., O.S. and A.N. reviewed and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: 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/08/80646) and was partially funded by the Federal Ministry of Education and Research within the Forschungscampus STIMULATE under grant numbers 13GW0473A and 13GW0473B.

Institutional Review Board Statement: Ethical review and approval were waived for this study because only retrospective analyses on publicly-available dataset were performed.

Informed Consent Statement: Not applicable as this study works with only publicly-available dataset.

Data Availability Statement: The datasets generated and/or analysed during the current study are available in the **COVID-19 image data collection** repository by *Joseph Paul Cohen*, https://github.com/ieee8023/covid-chestxray-dataset (Accessed on 30 May 2020) and in the **Chest X-ray Images (Pneumonia)** repository by *Paul Mooney*, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (Accessed on 30 May 2020).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [CrossRef]
2. Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K.S.; Lau, E.H.; Wong, J.Y.; et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N. Engl. J. Med.* **2020**, *382*, 1199–1207. [CrossRef]
3. Radiopaedia: COVID-19. Available online: https://radiopaedia.org/articles/covid-19-3 (accessed on 24 January 2024).
4. Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.; Sun, Z.; Xia, L. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. *Radiology* **2020**, *296*, E32–E40. [CrossRef]
5. Fang, Y.; Zhang, H.; Xie, J.; Lin, M.; Ying, L.; Pang, P.; Ji, W. Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. *Radiology* **2020**, *296*, E115–E117. [CrossRef]
6. Zhang, J.; Xie, Y.; Li, Y.; Shen, C.; Xia, Y. COVID-19 screening on chest x-ray images using deep learning based anomaly detection. *arXiv* **2020**, arXiv:2003.12338.
7. Narin, A.; Kaya, C.; Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks. *arXiv* **2020**, arXiv:2003.10849.
8. Apostolopoulos, I.D.; Mpesiana, T.A. COVID-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **2020**, *43*, 635–640. [CrossRef] [PubMed]
9. Kanne, J.P. Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: Key points for the radiologist. *Radiology* **2020**, *295*, 16–17. [CrossRef] [PubMed]
10. Bernheim, A.; Mei, X.; Huang, M.; Yang, Y.; Fayad, Z.A.; Zhang, N.; Diao, K.; Lin, B.; Zhu, X.; Li, K.; et al. Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection. *Radiology* **2020**, *295*, 685–691. [CrossRef] [PubMed]
11. Xie, X.; Zhong, Z.; Zhao, W.; Zheng, C.; Wang, F.; Liu, J. Chest CT for typical 2019-nCoV pneumonia: Relationship to negative RT-PCR testing. *Radiology* **2020**, *296*, E41–E45. [CrossRef] [PubMed]
12. Huang, P.; Liu, T.; Huang, L.; Liu, H.; Lei, M.; Xu, W.; Hu, X.; Chen, J.; Liu, B. Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion. *Radiology* **2020**, *295*, 22–23. [CrossRef]

13. Omer, S.B.; Malani, P.; Del Rio, C. The COVID-19 pandemic in the US: A clinical update. *JAMA* **2020**, *323*, 1767–1768. [CrossRef] [PubMed]
14. Rubin, G.D.; Ryerson, C.J.; Haramati, L.B.; Sverzellati, N.; Kanne, J.P.; Raoof, S.; Schluger, N.W.; Volpi, A.; Yim, J.J.; Martin, I.B.; et al. The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society. *Radiology* **2020**, *296*, 172–180. [CrossRef]
15. Harahwa, T.A.; Yau, T.H.L.; Lim-Cooke, M.S.; Al-Haddi, S.; Zeinah, M.; Harky, A. The optimal diagnostic methods for COVID-19. *Diagnosis* **2020**, *7*, 349–356. [CrossRef]
16. Jacobi, A.; Chung, M.; Bernheim, A.; Eber, C. Portable chest X-ray in coronavirus disease-19 (COVID-19): A pictorial review. *Clin. Imaging* **2020**, *64*, 35–42. [CrossRef] [PubMed]
17. Guan, W.J.; Ni, Z.Y.; Hu, Y.; Liang, W.H.; Ou, C.Q.; He, J.X.; Liu, L.; Shan, H.; Lei, C.l.; Hui, D.S.; et al. Clinical characteristics of coronavirus disease 2019 in China. *N. Engl. J. Med.* **2020**, *382*, 1708–1720. [CrossRef] [PubMed]
18. Durrani, M.; Inam ul Haq, U.K.; Yousaf, A. Chest X-rays findings in COVID 19 patients at a University Teaching Hospital—A descriptive study. *Pak. J. Med. Sci.* **2020**, *36*, S22. [CrossRef]
19. Wong, H.Y.F.; Lam, H.Y.S.; Fong, A.H.T.; Leung, S.T.; Chin, T.W.Y.; Lo, C.S.Y.; Lui, M.M.S.; Lee, J.C.Y.; Chiu, K.W.H.; Chung, T.; et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology* **2020**, *296*, E72–E78. [CrossRef]
20. Ng, M.Y.; Lee, E.Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M.M.s.; Lo, C.S.Y.; Leung, B.; Khong, P.L.; et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiol. Cardiothorac. Imaging* **2020**, *2*, e200034. [CrossRef]
21. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597.
22. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Acharya, U.R. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef]
23. Liu, J.; Cao, L.; Akin, O.; Tian, Y. Accurate and Robust Pulmonary Nodule Detection by 3D Feature Pyramid Network with Self-Supervised Feature Learning. *arXiv* **2019**, arXiv:1907.11704.
24. Yoo, S.; Gujrathi, I.; Haider, M.A.; Khalvati, F. Prostate cancer Detection using Deep convolutional neural networks. *Sci. Rep.* **2019**, *9*, 19518. [CrossRef]
25. Tô, T.D.; Lan, D.T.; Nguyen, T.T.H.; Nguyen, T.T.N.; Nguyen, H.P.; Phuong, L.; Nguyen, T.Z. Ensembled Skin Cancer Classification. ISIC 2019 Challenge Submission, 2019. Available online: https://hal.science/hal-02335240v1/file/Combined_approach_to_skin_cancer_classification.pdf (accessed on 24 January 2024) .
26. Vial, A.; Stirling, D.; Field, M.; Ros, M.; Ritz, C.; Carolan, M.; Holloway, L.; Miller, A.A. The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: A review. *Transl. Cancer Res.* **2018**, *7*, 803–816. [CrossRef]
27. Davenport, T.; Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **2019**, *6*, 94. [CrossRef] [PubMed]
28. Sloane, E.B.; Silva, R.J. Artificial intelligence in medical devices and clinical decision support systems. In *Clinical Engineering Handbook*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 556–568.
29. Mahadevaiah, G.; Rv, P.; Bermejo, I.; Jaffray, D.; Dekker, A.; Wee, L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Med. Phys.* **2020**, *47*, e228–e235. [CrossRef] [PubMed]
30. Agrebi, S.; Larbi, A. Use of artificial intelligence in infectious diseases. In *Artificial Intelligence in Precision Health*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 415–438.
31. Sweetlin, J.D.; Nehemiah, H.K.; Kannan, A. Computer aided diagnosis of drug sensitive pulmonary tuberculosis with cavities, consolidations and nodular manifestations on lung CT images. *Int. J. Bio Inspired Comput.* **2019**, *13*, 71–85. [CrossRef]
32. Yao, J.; Dwyer, A.; Summers, R.M.; Mollura, D.J. Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification. *Acad. Radiol.* **2011**, *18*, 306–314. [CrossRef] [PubMed]
33. Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. *Radiology* **2020**, *296*, E65–E71. [CrossRef] [PubMed]
34. Chen, J.; Wu, L.; Zhang, J.; Zhang, L.; Gong, D.; Zhao, Y.; Chen, Q.; Huang, S.; Yang, M.; Yang, X.; et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Sci. Rep.* **2020**, *10*, 1–11. [CrossRef]
35. Wang, L.; Wong, A.; Qui Lin, Z. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-ray Images. *arXiv* **2020**, arXiv:2003.09871.
36. Ghoshal, B.; Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. *arXiv* **2020**, arXiv:2003.10769.
37. Singh, G.; Yow, K.C. An interpretable deep learning model for COVID-19 detection with chest X-ray images. *IEEE Access* **2021**, *9*, 85198–85208. [CrossRef] [PubMed]
38. Singh, G.; Yow, K.C. Object or background: An interpretable deep learning model for COVID-19 detection from CT-scan images. *Diagnostics* **2021**, *11*, 1732. [CrossRef] [PubMed]
39. Shorten, C.; Khoshgoftaar, T.M.; Furht, B. Deep Learning applications for COVID-19. *J. Big Data* **2021**, *8*, 1–54. [CrossRef] [PubMed]

40. De Falco, I.; De Pietro, G.; Sannino, G. Classification of Covid-19 chest X-ray images by means of an interpretable evolutionary rule-based approach. *Neural Comput. Appl.* **2023**, *35*, 16061–16071. [CrossRef] [PubMed]
41. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
42. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
43. Mahendran, A.; Vedaldi, A. Salient deconvolutional networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 120–135.
44. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328.
45. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
47. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
48. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
49. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
50. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
51. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
52. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
53. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2818–2826.
54. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
55. Interpretable Machine Learning: A Guide for Making Black-Box Models Explainable. 2022. Available online: https://christophm.github.io/interpretable-ml-book (accessed on 24 January 2024).
56. Kopitar, L.; Cilar, L.; Kocbek, P.; Stiglic, G. Local vs. global interpretability of machine learning models in type 2 diabetes mellitus screening. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 108–119.
57. Kindermans, P.J.; Schütt, K.; Müller, K.R.; Dähne, S. Investigating the influence of noise and distractors on the interpretation of neural networks. *arXiv* **2016**, arXiv:1611.07270.
58. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv* **2014**, arXiv:1412.6806.
59. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; Volume 70, pp. 3145–3153.
60. Krug, A.; Knaebel, R.; Stober, S. Neuron Activation Profiles for Interpreting Convolutional Speech Recognition Models. In Proceedings of the NeurIPS Workshop IRASL: Interpretability and Robustness for Audio, Speech and Language, Montreal, QC, Canada, 8 December 2018.
61. Krug, A.; Ebrahimzadeh, M.; Alemann, J.; Johannsmeier, J.; Stober, S. Analyzing and visualizing deep neural networks for speech recognition with saliency-adjusted neuron activation profiles. *Electronics* **2021**, *10*, 1350. [CrossRef]
62. Krug, A.; Ratul, R.K.; Stober, S. Visualizing Deep Neural Networks with Topographic Activation Maps. *arXiv* **2022**, arXiv:2204.03528.
63. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2019; pp. 8024–8035.
64. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv* **2020**, arXiv:2009.07896.
65. Chatterjee, S.; Das, A.; Mandal, C.; Mukhopadhyay, B.; Vipinraj, M.; Shukla, A.; Nagaraja Rao, R.; Sarasaen, C.; Speck, O.; Nürnberger, A. TorchEsegeta: Framework for Interpretability and Explainability of Image-based Deep Learning Models. *Appl. Sci.* **2022**, *12*, 1834. [CrossRef]
66. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
67. PyTorch Reproducibility. Available online: https://pytorch.org/docs/stable/notes/randomness.html (accessed on 24 January 2024).
68. Nvidia Apex. Available online: https://github.com/NVIDIA/apex (accessed on 24 January 2024).

69. COVID-19 Image Data Collection. Available online: https://github.com/ieee8023/covid-chestxray-dataset (accessed on 24 January 2024).
70. Kermany, D.; Zhang, K.; Goldbaum, M. Labeled optical coherence tomography (oct) and chest X-ray images for classification. *Mendeley Data* **2018**, *2*, 651.
71. Chest X-ray Images (Pneumonia). Available online: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed on 24 January 2024).
72. Radiopaedia: Chest Radiograph. Available online: https://radiopaedia.org/articles/chest-radiograph?lang=us (accessed on 24 January 2024).
73. Diamond, M.; Peniston, H.L.; Sanghavi, D.; Mahapatra, S.; Doerr, C. *Acute Respiratory Distress Syndrome (Nursing)*; StatPearls Publishing: Treasure Island, FL, USA, 2021.
74. Matthay, M.A.; Zemans, R.L.; Zimmerman, G.A.; Arabi, Y.M.; Beitler, J.R.; Mercat, A.; Herridge, M.; Randolph, A.G.; Calfee, C.S. Acute respiratory distress syndrome. *Nat. Rev. Dis. Prim.* **2019**, *5*, 1–22. [CrossRef] [PubMed]
75. Fan, E.; Beitler, J.R.; Brochard, L.; Calfee, C.S.; Ferguson, N.D.; Slutsky, A.S.; Brodie, D. COVID-19-associated acute respiratory distress syndrome: is a different approach to management warranted? *Lancet Respir. Med.* **2020**, *8*, 816–821. [CrossRef] [PubMed]
76. Gattinoni, L.; Chiumello, D.; Rossi, S. COVID-19 pneumonia: ARDS or not? *Crit. Care* **2020**, *24*, 154. [CrossRef] [PubMed]
77. Bain, W.; Yang, H.; Shah, F.A.; Suber, T.; Drohan, C.; Al-Yousif, N.; DeSensi, R.S.; Bensen, N.; Schaefer, C.; Rosborough, B.R.; et al. COVID-19 versus non–COVID-19 acute respiratory distress syndrome: comparison of demographics, physiologic parameters, inflammatory biomarkers, and clinical outcomes. *Ann. Am. Thorac. Soc.* **2021**, *18*, 1202–1210. [CrossRef]
78. Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. IJDWM* **2007**, *3*, 1–13. [CrossRef]
79. Charte, F.; Rivera, A.J.; del Jesus, M.J.; Herrera, F. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing* **2015**, *163*, 3–16. [CrossRef]
80. Denise E. Morris, D.W.C.; Clarke, S.C. Secondary Bacterial Infections Associated with Influenza Pandemics. *Front. Microbiol.* **2017**, *8*, 1041.
81. Hanada, S.; Pirzadeh, M.; Carver, K.Y.; Deng, J.C. Respiratory Viral Infection-Induced Microbiome Alterations and Secondary Bacterial Pneumonia. *Front. Immunol.* **2018**, *9*, 2640. [CrossRef] [PubMed]
82. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
83. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 4765–4774.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.