EasyChair Preprint

№ 4494

Lightweight Deep Learning Models for Detecting COVID-19 from Chest X-ray Images

Stefanos Karakanis and Georgios Leontidis

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 18, 2022

Lightweight Deep Learning Models for Detecting COVID-19 from Chest X-ray Images

Stefanos Karakanis[a], Georgios Leontidis[a,*]

[a]*Department of Computing Science, University of Aberdeen, AB24 3UE, Aberdeen, UK*

Abstract

Deep learning methods have already enjoyed an unprecedented success in medical imaging problems. Similar success has been evidenced when it comes to the detection of COVID-19 from medical images, therefore deep learning approaches are considered good candidates for detecting this disease, in collaboration with radiologists and/or physicians. In this paper, we propose a new approach to detect COVID-19 via exploiting a conditional generative adversarial network to generate synthetic images for augmenting the limited amount of data available. Additionally, we propose two deep learning models following a lightweight architecture, commensurating with the overall amount of data available. Our experiments focused on both binary classification for COVID-19 vs Normal cases and multi-classification that includes a third class for bacterial pneumonia. Our models achieved a competitive performance compared to other studies in literature and also a ResNet8 model. Our binary model achieved 98.7% accuracy, 100% sensitivity and 98.3% specificity, while our three-class model achieved 98.3% accuracy, 99.3% sensitivity and 98.1% specificity. Moreover, via adopting a testing protocol proposed in literature, our models proved to be more robust and reliable in COVID-19 detection than a baseline ResNet8, making them good candidates for detecting COVID-19 from posteroanterior chest X-ray images.

Keywords: generative adversarial networks, deep neural networks, COVID-19, bacterial pneumonia, medical informatics, chest x-rays.

1. Introduction

Over the last couple of months, the world has been confronted by the rapid spread of COVID-19 resulting from the novel coronavirus, the SARS-CoV-2, which has already been named as one of the major events in the modern history. Part of the mitigation plans implemented across the globe has been the development of novel approaches to tackle the disease from a medical perspective, as well as to employ AI technologies to help to detect the disease from medical images, such as **C**hest **X**-**R**ay Images **(CXR)**.

This paper concerns the latter and specifically the development of a deep learning-based algorithm for accurate detection of Covid-19 from CXR. Arguably, various implementations have already been proposed

[*]Corresponding author

Email address: georgios.leontidis@abdn.ac.uk (Georgios Leontidis)

that use deep learning approaches as an attempt to address the detection of the COVID-19 from CXR. Deep learning algorithms are highly efficient on image acquisition which they can provide reliable results. [1] introduced a convolutional neural network namely COVID-Net to investigate how COVID-Net can make predictions using an explainability method, whereas [2] proposed three different convolutional neural networks models that achieved a 98% accuracy. Although more background information will be provided below, it is worth pointing out that one of the main limitations with the majority of the implementations proposed in literature is the limited amount of related COVID-19 data that are publicly available to use, which makes it significantly challenging to create generalisable models with overall high performance. Maguolo and Nanni [3] in their study presented a testing protocol that evaluates the bias of the model.

1.1. Research Contributions and Objectives

Our contributions can be outlined as follows:

- Considering that availability of CXR data for COVID-19 remains a challenge, one might need to resort to synthetic data to develop bespoke models that can generalise better to unseen examples, including different protocols. To this direction, in this paper we present a method for generating synthetic medical images using a previously presented deep learning Conditional Generative Adversarial Networks, adapted for our purpose (cGANs) [4]. Our main objective is to generate synthetic images to overcome the dataset limitation that lead to over-fitting.

- Implement two lightweight architectures that can detect the disease effectively with high accuracy and robustness, and compare them with other state of the art models presented in literature.

- we follow the *testing protocol* proposed by Maguolo and Nanni [3] for evaluating our models in COVID-19 detection, which we find it to be important for the reliability of results.

Specifically, in this paper we implemented two deep neural networks; the first for detecting COVID-19 vs normal cases (no disease present), with the second targeting three different cases, namely **bacterial pneumonia**, **COVID-19** and **normal**. Ultimately, we make comparison of our models with other state-of-the art models presented in literature.

2. Background

Machine Learning technologies have seen an unprecedented success and uptake in the last decade or so, with state-of-the-art results achieved across various application areas that span generic computer vision tasks [5, 6, 7], medical imaging and geometry [8, 9, 10, 11, 12], natural language processing [13, 14] and signal processing [15, 16], to name a few. Carrying on from this it was only a matter of time before similar success was evidenced on implementing machine learning technologies to early diagnose patients with COVID-19, employing methods that can process and interpret medical imaging data, such as X-ray images and computed tomography (CT) scans. An example of such technology is being experimentally used in hospitals to screen mild cases, triage new infections, and monitor the progression of the disease [17] through the use of deep learning techniques.

However, numerous studies have shown that such AI tools perform inadequately and come with various limitations. [17] mention in their study " *This review indicates that proposed models are poorly reported, at high risk of bias, and their reported performance is probably optimistic.*" Additionally, one of the biggest limitations of implementing a prediction model for medical image detection and especially for COVID-19 disease is data availability.

Being aware of these constrains, in this paper we adopt the protocol mentioned by Maguolo and Nanni [3] and take the advantage of state-of-the-art techniques in order to overcome the above mentioned limitations and perform detection more effectively and reliably.

2.1. Literature Review

There exists a very large dataset with chest X-rays for pneumonia disease, released by Kermani et al. [18]. On the other hand, information for COVID-19 has been very limited and the amount of images corresponding to this virus is limited. Therefore, developing robust techniques to detect COVID-19 from CXR remains an open challenge; nevertheless a comprehensive body of literature has already been proposed, which we aim at summarising below.

2.1.1. Transfer Learning

An established method to address the issue of limited data has been the use of transfer learning approaches, which can improve the learning process in a new problem through the transfer of knowledge from a related problem that has already been learned and solved.

Narin et al. [2] constructed a small dataset of 50 COVID-19 cases collected from Cohen et al. [19] repository and 50 normal cases extracted from Kaggle [20]. They utilised a five fold cross validation, a re-sampling method that evaluates machine learning models on a limited data sample, to train and test a ResNet50 [21] model, achieving an accuracy of 98%.

Castiglioni et al. [22] presented an entirely different approach and collecting a dataset that was not made public. Their train set included 250 COVID-19 cases and 250 normal (no disease present) images while the test set comprised 74 COVID-19 cases and 36 normal ones unrelated from the train set. Various 10 ResNet models [21] were trained on this dataset and obtained a ROC-AUC of 0.80 for the classification task. The performance of the model was much lower in comparison to the other studies reported in literature. Moreover, the authors applied classification in both CXR projections (e.g anteroposterior and posteroanterior), which makes it more robust according to the critical appraisal of Maguolo and Nanni testing protocol [3].

Soares et al. [23] used convolutional neural networks to detect COVID-19 cases. The dataset comprised 175 COVID-19, 100 normal and 100 pneumonia annotated CXR images. Moreover, they utilised transfer learning technique with ImageNet [24] on three different architectures, Xception [25], ResNet [21], and VGG-16 [26]. Their results show that all models performed well with high accuracy, especially the VGG-16 model. However, upon reflection they observed that their models require further improvement, suggesting as future directions to " *evaluate models with different architectures, parameters, and datasets that use augmentation techniques.*"

Wang and Wong [1] introduced a new architecture named Covid-Net. They used transfer learning [27] on ImageNet [24] as well. Furthermore, they created a large dataset with 183 COVID-19 cases, 5,538 with Pneumonia and 8,066 normal ones. They produced a test set of 100 images with pneumonia and normal lungs but only 31 of COVID-19 cases. In their work, they made explicit that there is no overlap between the test and the training set of the patients. That is of great importance in tasks of this manner. Wang and Wong [1] mentioned that Covid-Net achieved 92% accuracy.

2.1.2. Augmentation and Generation Techniques

A study presented in [28] described an efficient multi-class deep learning method for detecting COVID-19 vs normal vs pneumonia from CXR. They collected a total of 295 images of COVID-19. Additionally, they collected 65 examples of normal cases as well as 98 cases showing pneumonia. With the purpose of overcoming the limitation of the COVID-19 dataset, they restructured the data classes of the dataset by employing a Fuzzy Color technique as a pre-processing step and the images that were organised together with the original ones were stacked. In their results they demonstrated 100% success in the classification of COVID-19 images, and 99.27% success for the classification of Normal and Pneumonia images.

Generative Adversarial Networks (GANs) have demonstrated remarkable performance in various settings, including healthcare. The main idea in training a GAN is in the form of a zero-sum game, in which one network tries to discriminate between real and synthetic images, with the other one - generator - trying to fool the discriminator by producing artificial images that the discriminator considers as real, therefore creating images similar to real ones.

Loey et al. [29] motivated from the lack of the data, utilised GAN architecture [30] to synthesise auxiliary images to support in the detection of this disease from the available CXR images in the interest of achieving high performance. Additionally, they employed three different deep learning models with transfer learning [27]. Initially, the dataset comprised 306 CXR images across four categories, i.e. COVID-19, normal, bacterial pneumonia and viral pneumonia. By employing a GAN architecture, they increased the dataset images to be 30 times larger than the original set. That led to a total of 8100 images across four classes. The training set consisted of the 70% of the data, with validation and test sets the rest 20% and 10% respectively. The authors mentioned that they achieved 100% in testing accuracy and 99.9% in the validation accuracy.

In table 1 we present the summary of the aforementioned models.

Table 1: Summary of state of the art deep learning models for COVID-19 detection

Literature	Number of Cases	Architecture	Accuracy
Narin et al. [2]	50 COVID-19, 50 Normal	ResNet50	98.0%
Castiglioni et al. [22]	250 COVID-19 , 250 normal images	10 ResNets	80.0%
Soares et al. [23]	175 COVID-19, 100 normal, 100 pneumonia	VGG-16	97.3%
Wang et al. [1]	183 COVID-19, 5,538 pneumonia , 8,066 normal	CovidNet	92.0%
Toğaçar et al. [28]	295 COVID-19, 65 normal, 50 pneumonia	MobileNet V2 [31]	97.06%
Loey et al. [29]	69 COVID-19, 79 Normal, 79 Bacterial, 79 Viruses	GoogLeNet [32]	100%

2.2. Dataset Evaluation

As previously mentioned, Maguolo and Nanni [3] made an evaluation of numerous of models that are primarily developed for COVID-19 auto detection. Their evaluation was based on the testing protocols of those models' detection performance. In addition, it was mentioned that those testing protocols are inequitable thereby, the neural networks are learning patterns in the dataset which are not correlated to the presence of COVID-19. The results of their study shown that "*these protocols might be biased and learn to predict features that depend more on the source dataset than they do on the relevant medical information*" [3].

Regardless the method that has been used in pursuit of detecting COVID-19, none of all the aforementioned techniques used such a testing protocol to validate the model learning patterns in the dataset. Therefore, [3] recommended that proposed solutions have to adopt a testing protocol and methodology to evaluate how bias they are when it comes to classifying the images based on the presence of the disease or the underlying characteristics of the image, i.e. capturing device and/or protocols, given that the released dataset of COVID-19 CXR is a product of aggregating data across several hospitals.

2.3. Description of Datasets

In this study, we utilise datasets that are publicly available as mentioned above. We evaluate our proposed methods in two settings, one considering the detection of COVID-19 vs Normal cases with the other one focusing on a three-class problem, adding a third class, namely pneumonia. The COVID-19 dataset has been made available to the community via a GitHub repository by a researcher named Joseph Paul [19] based at the Montreal University. This dataset is the main source of data across most of the papers addressing

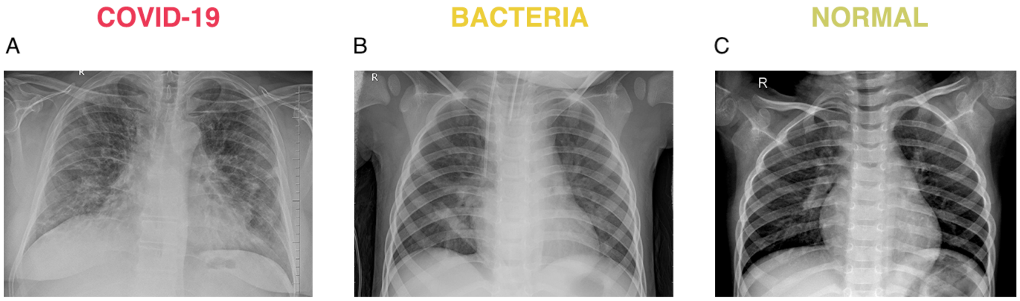

Figure 1: Illustration of the classes from both datasets [19] & [20].

COVID-19 detection and is being regularly updated with new images shared by researchers from different domains.

The second dataset consists of normal CXR cases and pneumonia CXR cases and was obtained from a Kaggle [20] that includes 5856 grey-scaled images, containing train, validation and test sets.

In figure 1 one can see examples of the data used in our studies. The annotation with **'A'** is obtained from the COVID-19 [19] while the **'B'** and **'C'** extracted from the second dataset by [20].

3. Methodology

3.1. Dataset

For the implementations presented in this paper we utilised all the available COVID-19 CXR in posteroanterior (PA) chest view (145 images). The Kaggle [20] dataset is structured into two folders (normal, pneumonia) and contains sub-folders for each image set (train/val/test). Considering the large amount of data found in the two classes, normal and pneumonia, we randomly extracted the same amount of images as in the COVID-19 cases so that we end up with a balanced dataset. Considering our experiments concern the use of synthetic images and also comparing the performance against models that have not made use of synthetic images, our datasets have as follows:

- Without Synthetic Images: 145 with COVID-19, 145 with Bacterial Pneumonia and 145 Normal
- With Synthetic Images: 275 with COVID-19 (130 synthetic), 275 with Bacterial Pneumonia (only real images) and 270 Normal (only real images)

3.2. Image data augmentation

We utilise image data augmentation techniques on the training set to mitigate the dataset limitation issue. Such an approach has been found to improve a model's performance and generalisation capabilities when tested against new images. The purpose of this process is to create new plausible versions of CXR that can help to improve the representation learning process via having more examples in place during training. Besides in this specific problem we are tackling, it is not uncommon CXR to be misaligned or slightly rotated.

3.3. Image synthesis

A second approach to address the problem of low data (and overfitting) resulting from the limited number of images in the dataset, can be achieved by utilising a cGAN [4] architecture. As will be shown later, via employing such an approach we managed to generate realistic synthetic images, which helped to improve the performance of our models extensively (Figure 3).

Across the board, we adopted a cGAN architecture and performed fine-tuning for synthesising high quality CXR for the Covid-19 class. The reason is that there exist a plethora of normal and bacterial pneumonia CXR images, hence it would be unnecessary to generate synthetic images for those classes. Initially, we performed image pre-processing to ensure that all CXR images have equal size and aspect ratio. Following extensive experimentation and tuning of our cGan model, we present the setup of the architecture (discriminator, generator) used in this study, that led to a considerable increase in performance.

Discriminator: The discriminator network has a standard CNN formation that receives the input image of size 446x446x1 (lesion ROI), and produces a binary decision: whether the image is real or fake. The network is composed of four dense layers and a final output layer. Drop out layers are also applied after every dense layer. Embedded layer is applied as a lookup table to map from integer indices to dense vectors and then is flattened to be equal to the amount of elements contained in tensor. Lastly, Leaky ReLU activation functions are employed to all layers apart from the output layer which utilises the Sigmoid function for the likelihood of a range between 0 and 1 [0,1] score of the image.

Generator: The generator network receives a vector of 100 random numbers (latent space) pulled from a uniform distribution as input and outputs a lungs lesion image of size 446x446x1. The network architecture consists of a fully connected layer reshaped to size 4x4x128 and four dense neurons to up-sample the image. Convolution over the up-sampled image provides a larger output image. Batch-normalization is applied to each layer of the network with momentum of 0.8 value, except for the output layer. Normalising the outputs of each neuron as an effect of zero mean and unit variance through the entire mini-batch levels off the cGAN learning process and avoids the generator from collapsing all samples to a singular point [30]. Nevertheless, GANs and its variations cannot be in an absolute convergence. ReLU activation functions are used to all layers except the output layer which utilises the hyperbolic tangent (tanh) activation function and inputs to the generator and discriminator are scaled to the range between -1 and 1 [-1,1].

We trained the cGAN by setting the slope of the curve of the Leaky ReLU to $alpha = 0.2$. Furthermore weights were initiated to a zero-centered normal distribution. We employed Adam optimizer for stochastic gradient descent [33], an adaptive moment estimation that embeds the initial and second moments of the gradients, handled by parameters $p = 0.5$ and learning rate of 0.0002 for 4000 epochs.

3.3.1. ***Evaluation***

In terms of evaluating the cGAN model, we conducted research of couple of qualitative and quantitative techniques. This is due to the fact that, it is not possible to objectively assess the progress of the training or to measure the quality of the model from the loss function alone. Hence, we visually examined the resulting

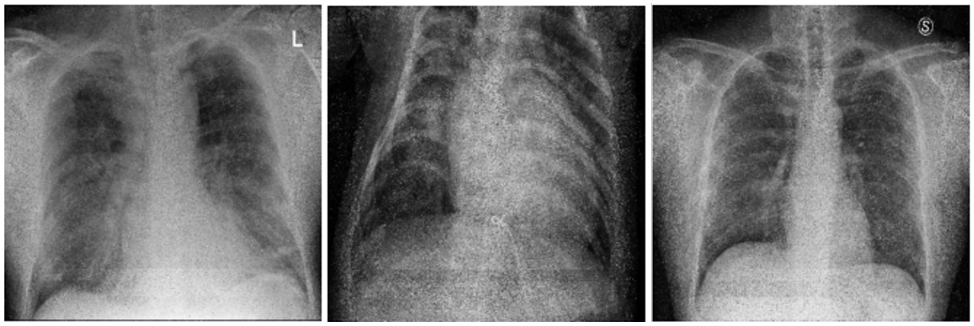

Figure 2: cGAN Generated COVID-19 samples after 3230 epochs.

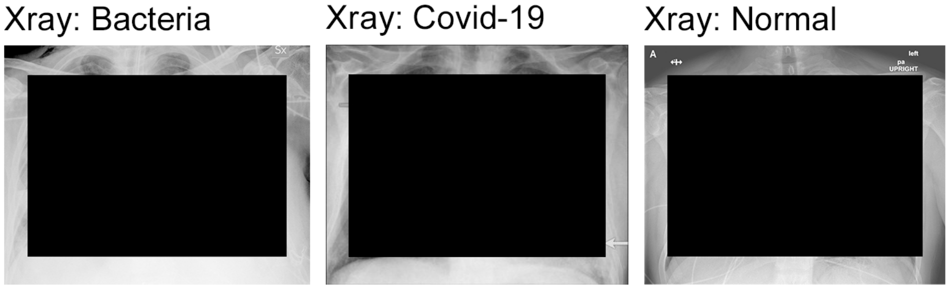

Figure 3: A sample of cases pertaining to bacterial pneumonia, COVID-19 and Normal conditions from the mask dataset.

images, that is, a human expert assessed the quality of the images compared to the distribution of examples found in real images.

During the training process, various models were saved systematically across training epochs (e.g. 50, 100, 200) and generating a sample image of the three cases (COVID-19, Bacterial Pneumonia and Normal) to observe the output of the generator along with the performance plot, as there is no objective measure of model performance. This enables for the post-hoc evaluation of every saved generated model from its generated sample images.

Figure 2 show examples of generated images after 3230 epochs. All the images have the size of 446x446, which is the default output size of our cGAN architecture.

3.4. Masked Images

Aiming at testing our model bias in the dataset used based on the study by Maguolo and Nanni [3], we created a second version of our entire dataset, but with a different image pre-processing technique. The variation in this dataset is that the center of each image contains a black rectangle shape object in such way that removes the anatomical structure in the chest and lungs. Figure 3 provides an example of how the masked dataset looks like.

3.5. Models

For the detection process in the binary classification setting, we employed one of the most popular – relatively shallow – architectures for image detection: ResNet8 [21] appended with small modification and

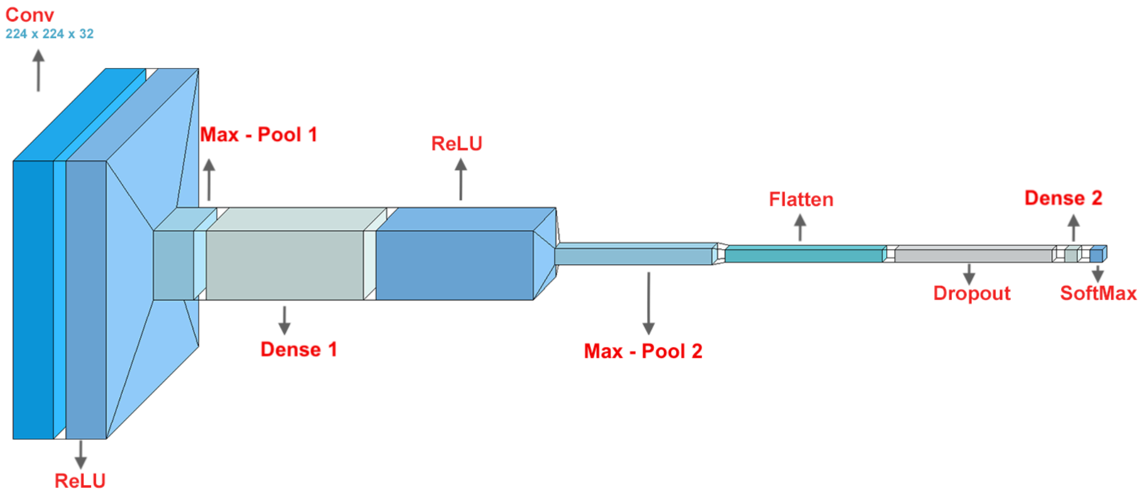

Figure 4: Architecture of the proposed binary model.

we compared it with our proposed model. Furthermore, we proposed an additional approach in a multi-class setting, which performs detection across the three conditions we are focusing on: COVID-19 vs bacterial pneumonia vs normal, and it is compared with other implementations proposed in literature.

3.5.1. ResNet8

In binary classification we adopted a popular lightweight architecture ResNet8 [21] pre-trained in the defacto ImageNet dataset [24]. This architecture is selected as a base model due to its small size and reflects the decreased complexity, memory consumption and duration.

In our case we used CNN structure that receives input image of size 224 x 224 x 3 and outputs a binary decision COVID-19 or Normal. The network consists of seven convolutional layers with filter size of 64, kernel 3 x 3 and ReLU for activation function. At the end of the convolutional layers we used global average pooling [34] for minimising overfitting. Finally, a dense layer along with ReLU is added followed by a dropout layer.

3.5.2. Proposed models

For the binary classification case (Figure 4), we propose a deep convolutional neural network (CNN) for the detection of COVID-19. The model has been designed in a form to support a lightweight architecture without transfer learning, whilst performing well. It can deal with any non-uniformity in the data distribution and the limited accessibility of training images in the classes. It consists of a single convolutional layer with filter size 32 and kernel 4x4, followed by ReLU [35] activation function and Max Pooling layer for down-sampling the image (input representation) and enabling feature extraction. After a flatten layer there exists a dense layer of size 128, followed by dropout and a final dense layer with softmax activation function for a binary output. The total number of trainable parameters of our proposed binary model is 49,058. For comparison the ResNet8 model used has 221,186. Similarly to the binary classification case, our multi-class architecture (Figure 5)

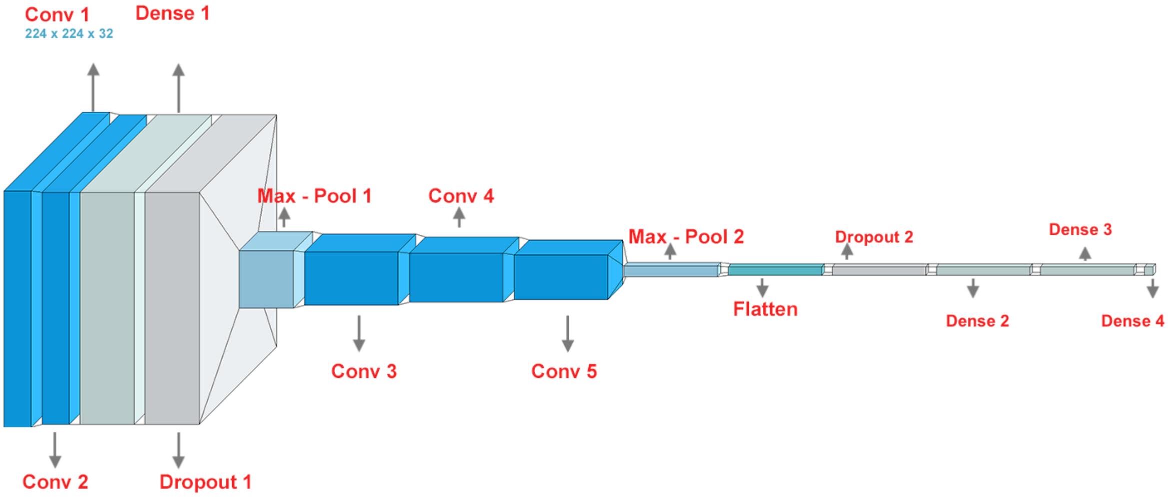

Figure 5: Architecture of the proposed multi-class model.

follows analogous lightweight design, albeit adapted to accommodate three classes. The architecture consists of five Convolutional layers, all with ReLU activation function. The first convolutional layer uses two strides, a filter size of 32 and kernel 4x4, while the second one uses one stride to perform convolution with 64 filters and 4x4 kernel. After performing a max pooling operation, three additional convolutional layers are added, with 128, 128 and 256 filters respectively and same kernel size. This is then followed by a max pooling layer and a flatten layer. Finally, the model includes two dense layers with ReLU activation function (512 and 128 size respectively) and the final dense layer with softmax activation function for the three-class output. The total number of parameters of our proposed multi-class model - excluding the Grad-CAM layers - is 87,171.

3.5.3. Training process

First phase involves the pre-processing method which primarily concerns the augmentation and artificial data generation processes. The former, refers to the creation of new image variations such as flip and rotate while the latter to generating new images by using Conditional Generative Adversarial Networks (cGAN). Second phase concerns the training and hyperparametrisation process on both approaches adopted in this study, i.e. masked images and unmasked images. The ResNet8 model utilises pre-trained weights that are loaded from the ImageNet [24] dataset over the transfer learning method. The Proposed models do not use transfer learning technique as they are prioritising lightweight architecture. For the binary classification model, categorical_crossentropy and Adam are used as loss and optimisation functions respectively. Furthermore, loss function for both ResNet8 and proposed multi-class model is the categorical cross entropy.

The CXR images are used to train the models with certain size of 224 x 224 pixels and three channels of colour (RGB). In training process it is used 20% of the images for validation/test data and the rest 80% for training data. The results presented below are organised in a confusion matrix and Tables, and are also

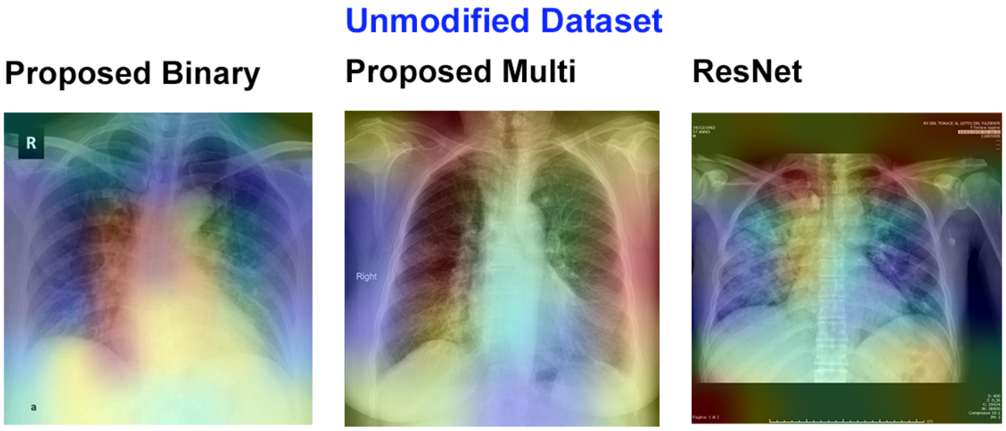

Figure 6: Grad-CAM heatmap on three models with COVID-19 unmodified images

examined using a class activation map (Grad-CAM heatmap [23]) for visual inspection. The models were developed and trained on Google Co-laboratory Virtual Environment utilising TensorFlow [36], Keras [37] and visualisation Python libraries. The implementation proposed in this paper will be openly released via a github repository.

4. Models Evaluation

Considering that most of the COVID-19 implementations were performed with inadequate evaluation as very nicely described in [3], we decided to adopt the principles outlined in this study and introduce several evaluation approaches in order to provide more reliable results and strengthen findings. We generated confusion matrices and employed the Grad-CAM method as well that provides a visual illustration on the areas that contributed to the models' outcome. Ultimately, we corroborate that generated medical images can be employed for synthetic data augmentation, contributing to an increased performance of CNN models for medical image classification in the presence of limited data.

4.1. Grad-CAM heatmap

In our evaluation we present results of our models' detection performance, whereby Grad-CAM [38] approach is used as a means to provide visual explanation of the decision. As shown in figure 6, both our proposed models highlighted as more important region (red color shades) the area around the lungs; our multi-class model also demonstrated that some areas around thorax are important for the decision making process.

By employing the testing protocol outlined in [3], our binary model demonstrated very poor and random results on the actual image by highlighting random regions on the image. Similarly, our proposed multi-class model showed interest in random regions on the image. On the other hand, ResNet8 demonstrated interest in the left side of the image, which according to the study by Maguolo and Nanni, it indicates bias on the

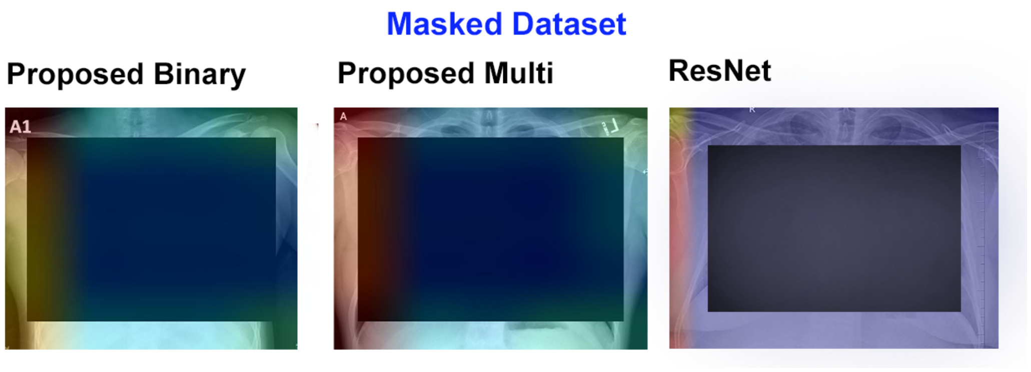

Figure 7: Grad-CAM heatmap on three models with COVID-19 masked images.

COVID-19 dataset, something that our proposed models have been shown to be more robust. Figure 7 shows an example of Grad-CAM on masked images.

4.2. Confusion Matrix

With the intention to provide more information on the performance of our classifiers on test images we provide a series of confusion matrices below. Such a matrix shows the number of true and false predictions and how they are distributed across each class and scenario.

Looking at figure 8 it is evident that our proposed binary model identified all of the 59 real COVID-19 images correctly while only 3 out of 59 normal cases were misidentified as COVID-19. This shows that our model is highly accurate in COVID-19 detection. Similarly, our proposed multi-class model for bacterial pneumonia vs COVID-19 vs normal cases indicates that in the bacterial pneumonia cases 97% of them were classified correctly with only a 3% percent being misclassified as normal. Across the COVID-19 cases only 1% were wrongly detected as normal, while the 99% of them was classified correctly. Along the same lines, 98% of the normal cases were identified properly and just 2% of them were misclassified as bacterial pneumonia. It is Worth mentioning that our proposed multi-class model detected with 99% accuracy the cases of COVID-19 from the test set, which is of significant importance in a medical disease detection context. Furthermore, in ResNet8 it is apparent that from the total of 59 COVID-19 CXR images, the 47 were correctly classified as COVID-19 while the remaining 12 were misclassified as normal. On the other hand, all the normal cases were classified correctly. All the above numbers and the ones found in 2 refer to the dataset that included real and synthetic images – about double the size of the one with only real images. Nevertheless, the performance difference has been about 2% for the binary classification setting and 4% for the multi-class classification, both in favour of the models augmented with synthetic images. It is worth clarifying that for the sake of fairness, the test set (real images) has been the same in both synthetic / real and real-only settings.

We perform similar to [3] technique to assess our models' performance in classifying the different conditions, while the region with the particular anatomical information being obscured. In figure 9 we illustrate the performance on this testing protocol. It is clear that our proposed binary model performs poorly across

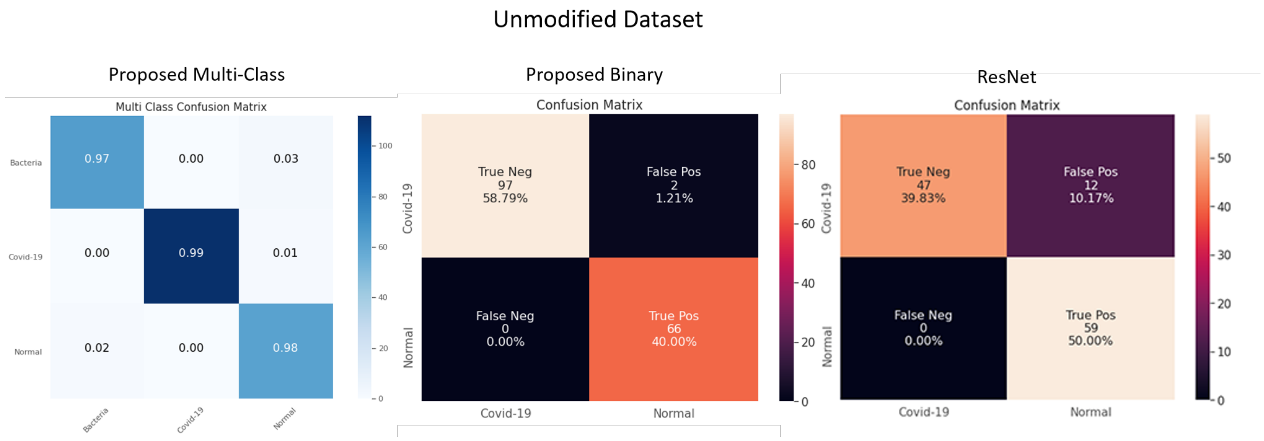

Figure 8: Confusion Matrix of our three models for unmodified test set.

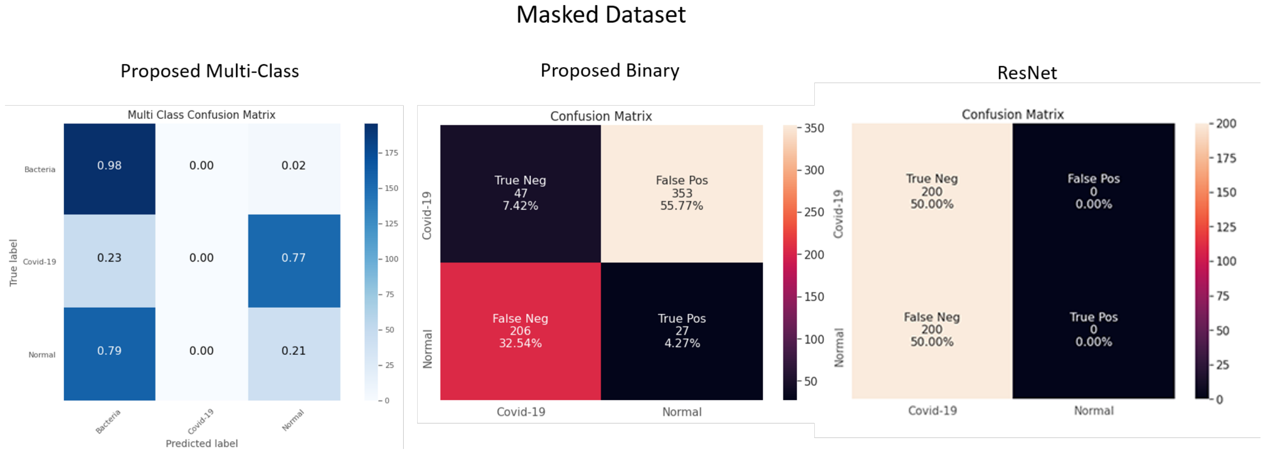

Figure 9: Confusion Matrix of our three models for masked test set.

the whole COVID-19 CXR dataset, as the model classified correctly just 47 images. Similarly, our proposed multi-class model presents a similar trend. For instance, in COVID-19 images the model was not able to detect any COVID-19 cases. In contrast, the model classified 23% of them wrongly as bacterial pneumonia while the remaining 77% were detected incorrectly as normal. Furthermore, 21% of the normal cases were identified properly, with 79% misclassified as bacterial pneumonia. By comparison, ResNet8 model demonstrated that from the whole COVID-19 CXR dataset, all of them were correctly classified as COVID-19. In addition, all normal cases were also classified as COVID-19 ones.

It is apparent that both of our proposed models – in contrast to the pre-trained ResNet8 model – were not able to identify the majority of COVID-19 images correctly when following the testing protocol with masked images. Therefore, it can be inferred that our models are not biased in favour of COVID-19 dataset, which is one of the main strengths of this study.

Table 2: Comparing performance across ResNet8, ResNet16 and proposed models with and without synthetic Covid-19 images

Metric	ResNet8	ResNet16	P. Binary	P. Binary with cGAN	P. Multi	P. Multi with cGAN
Unmodified Dataset						
Accuracy	89.8%	93.1%	96.5%	98.7%	94.3%	98.3%
Specificity	100%	97.4%	94.2%	98.3%	93.1%	98.1%
Sensitivity	76.2%	89.8%	95.3%	100%	95.6%	99.3%
Masked Dataset						
Accuracy	50%	50%	14.2%	13.5%	45.0%	42.9%
Specificity	100%	100%	24.5%	22.1%	34.4%	31.1%
Sensitivity	0.0%	0.0%	4.0%	3.9%	12.3%	11.2%

4.3. Discussion on Results

Table 2 presents an overall comparison of the models on each of the two evaluation settings, namely the unmodified dataset and also the masked one. It can be observed that the highest performance overall in binary classification was achieved by our proposed model, which achieved a 98.7% accuracy and 100% sensitivity, which means that our model predicted all COVID-19 cases correctly. In addition, specificity was very high as well at 98.3%, meaning that all normal cases were identified correctly. We believe this performance to be highly desirable in a healthcare setting, because it is more preferable to misclassify normal cases as COVID-19 – whereby the person will undergo further examinations – than missing out COVID-19 cases. In essence, if the sensitivity of our model had been lower, our model would have misclassified COVID-19 cases as normal and the patient would not have been considered for additional testing, which is not desirable. In addition, our multi-classification model performed well too, with accuracy of 98.3%. Moreover, sensitivity and specificity were very high at 99.3% and 98.1% respectively, which makes it very robust and reliable. Furthermore, ResNet8 performed well in detecting normal but performed poorly in identifying COVID-19 cases.

Finally, our proposed models demonstrated a very poor and random performance – as expected – when it came to classifying the masked dataset. By comparison, ResNet8 classified the entire set of COVID-19 correctly which implies, as reported earlier and in literature, that some pretrained models might be biased and prone to overfitting the COVID-19 dataset, by learning features that correspond to the image itself rather than the anatomical structure they represent (see Table 2). This needs to be considered along with Figures 8 and 9.

4.4. Comparison with state-of-the-art models

In this last section, we make comparisons between pre-trained deep neural networks that have been proposed for detecting COVID-19 to date against our proposed models in binary and multi-classification

Table 3: Comparison of state of the art models with our cGAN-based proposed models

Literature	Subjects	Task	Method	Accuracy
Proposed (binary)	275 COVID-19, 275 Normal	Detection	CNN	98.7%
Proposed (multi-class)	275 COVID-19, 275 Bacteria, 275 Normal	Detection	CNN	98.3%
Resnet18 [39]	624 Images for Normal and COVID-19	Image Generation and Detection	ResNet18, GAN	99.0%
COVIDx [40]	45 COVID-19, 1203 Normal, 931 Bacterial, 660 Viral	Detection	ResNet-50	96.23%
Narin et al. [2]	50 COVID-19, 50 Normal	Detection	ResNet	98.0%
COVIDNet [41]	183 COVID-19, 551 Pneumonia, 8066 Normal	Detection	COVIDNet-CXR Small and COVID-Net-CXR Large	92.6%

settings.

In table 3 we are showing results from some of the latest approaches that have been proposed for COVID-19 detection. The highest score was achieved by a ResNet18 model [39], which demonstrated a 99% accuracy in a binary classification setting for Normal and Pneumonia conditions. The most interesting component in this study is that they synthesised new images in order to improve the performance of their pre-trained model. However, this score refers only to those two cases excluding COVID-19. On the other hand, Narin et al. [2] modified a pre-trained ResNet which was trained on only 50 images of COVID-19 and 50 of Normal images. They acquired 98% accuracy which was compared to other models (i.e. AlexNet). Additionally, COVIDx [40] demonstrated a similar performance of 96% achieved with a pre-trained ResNet50 model in a four-class setting, including a set of 45 COVID-19 CXR images.

By comparison, our models perform similar to the aforementioned models but are evaluated on the testing protocol in [3] as well, along with the standard evaluation process that all the other studies have employed that is common in machine learning settings. Moreover, we did not use any pre-trained weights to our models but only real and synthetic/artificial images from the cGAN method we developed, primarily because of the nature of medical images which are distinctly different to the images found in the ImageNet database. Therefore, we consider both models to be competitive with respect to the state of the art models for COVID-19 detection, but being more robust and reliable given the results we demonstrated using the testing protocol in [3]. Besides, a ResNet18 model used in previous studies has a considerably higher number of trainable parameters (11M [42]) than our proposed models.

5. Conclusion and Future Work

Arguably, this current pandemic has transformed our lives to an unprecedented extent. However, the effort of the research community has been immense across various fronts and to this direction this paper proposed a simple approach for reliably detecting COVID-19 across various scenarios. We showed how robust our method is via experimenting with a masked dataset based on the protocol proposed by Maguolo and Nanni [3], demonstrating that our approach learns proper features and not features pertaining to the image protocol itself or other irrelevant information, which would denote a biased model. In addition, via employing a GAN [30] approach we were able to improve the performance of our approaches, given the very limited amount of real data available.

From our results, we demonstrated that simple models, such as our proposed models for binary and multi-classification, in conjunction with conditional generated adversarial networks (cGANs) for synthetic image generation, can achieve high performance and accuracy in COVID-19 detection, without requiring to utilise pre-trained weights in the model, which can increase a model's size, parameters and complexity. Our proposed (binary) and proposed (multi) demonstrated no bias in COVID-19 detection and therefore are the optimal deep learning networks for detection of COVID-19 in posteroanterior CXR images. In conclusion, rapid, accurate and accessible tools are required to assist detection and management of COVID-19 from CXR. Furthermore, adaptation of new testing protocols on new or existing models might be recommended. We hope that in the near future we will be able to improve our techniques and propose new ones, as more real data become available. Relying solely on synthetic data might not be the way forward as capturing all the variability found in medical images might require larger amounts of real data.

Acknowledgment

The authors would like to thank the multiple teams that have contributed to the release of the datasets used in this paper. We would also like to thank the Data Lab, which provided an MSc AI scholarship to the first author, making this project possible.

References

[1] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images. *arXiv preprint arXiv:2003.09871*, 2020.

[2] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

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

[4] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[5] Fabio De Sousa Ribeiro, Francesco Calivá, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis, and Stefanos Kollias. Deep bayesian self-training. *Neural Computing and Applications*, pages 1–17, 2019.

[6] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020.

[7] Fabio De Sousa Ribeiro, Georgios Leontidis, and Stefanos Kollias. Introducing routing uncertainty in capsule networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[8] Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[9] Berkman Sahiner, Aria Pezeshk, Lubomir M Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H Cha, Ronald M Summers, and Maryellen L Giger. Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36, 2019.

[10] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, page 104065, 2020.

[11] Georgios Leontidis, Bashir Al-Diri, and Andrew Hunter. A new unified framework for the early detection of the progression to diabetic retinopathy from fundus images. *Computers in biology and medicine*, 90:98–115, 2017.

[12] Georgios Leontidis, Bashir Al-Diri, and Andrew Hunter. Diabetic retinopathy: current and future methods for early screening from a retinal hemodynamic and geometric approach. *Expert Review of Ophthalmology*, 9(5):431–442, 2014.

[13] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[14] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[15] Francesco Caliva, Fabio Sousa De Ribeiro, Antonios Mylonakis, Christophe Demaziere, Paolo Vinai, Georgios Leontidis, and Stefanos Kollias. A deep learning approach to anomaly detection in nuclear reactors. In *2018 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2018.

[16] Danfeng Xie, Lei Zhang, and Li Bai. Deep learning in visual computing and signal processing. *Applied Computational Intelligence and Soft Computing*, 2017, 2017.

[17] Laure Wynants, Ben Van Calster, Marc MJ Bonten, Gary S Collins, Thomas PA Debray, Maarten De Vos, Maria C Haller, Georg Heinze, Karel GM Moons, Richard D Riley, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *bmj*, 369, 2020.

[18] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[19] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020.

[20] Kaggle. chest x-ray images(pneumonia) 2020, kaggle, 2020.

[21] 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*, pages 770–778, 2016.

[22] Isabella Castiglioni, Davide Ippolito, Matteo Interlenghi, Caterina Beatrice Monti, Christian Salvatore, Simone Schiaffino, Annalisa Polidori, Davide Gandola, Cristina Messa, and Francesco Sardanelli. Artificial intelligence applied on chest x-ray can aid in the diagnosis of covid-19 infection: a first experience from lombardy, italy. *medRxiv*, 2020.

[23] Lucas P Soares and Cesar P Soares. Automatic detection of covid-19 cases on x-ray images using convolutional neural networks. *arXiv preprint arXiv:2007.05494*, 2020.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[28] Mesut Toğaçar, Burhan Ergen, and Zafer Cömert. Covid-19 detection using deep learning models to exploit social mimic optimization and structured chest x-ray images using fuzzy color and stacking approaches. *Computers in Biology and Medicine*, page 103805, 2020.

[29] Mohamed Loey, Florentin Smarandache, and Nour Eldeen M Khalifa. Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4):651, 2020.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[32] 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*, pages 1–9, 2015.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Norman L Strominger, Robert J Demarest, and Lois B Laemle. *Noback's human nervous system: structure and function*. Springer Science & Business Media, 2012.

[35] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[36] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[37] François Chollet et al. Keras. `https://keras.io`, 2015.

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

[39] Nour Eldeen M Khalifa, Mohamed Hamed N Taha, Aboul Ella Hassanien, and Sally Elghamrawy. Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and

a fine-tuned deep transfer learning model using chest x-ray dataset. *arXiv preprint arXiv:2004.01184*, 2020.

[40] Muhammad Farooq and Abdul Hafeez. Covid-resnet: A deep learning framework for screening of covid19 from radiographs. *arXiv preprint arXiv:2003.14395*, 2020.

[41] Hokuto Hirano, Kazuki Koga, and Kazuhiro Takemoto. Vulnerability of deep neural networks for detecting covid-19 cases from chest x-ray images to universal adversarial attacks. *arXiv preprint arXiv:2005.11061*, 2020.

[42] Mei Chee Leong, Dilip K Prasad, Yong Tsui Lee, and Feng Lin. Semi-cnn architecture for effective spatio-temporal learning in action recognition. *Applied Sciences*, 10(2):557, 2020.