Covid4HPC: A Fast and Accurate Solution for Covid Detection in the Cloud using X-Rays*

Dimitrios Danopoulos[1 [0000−0001−9327−5983]], Christoforos Kachris[1,2 [0000−0003−0818−1902]], and Dimitrios Soudris[1 [0000−0002−6930−6847]]

[1] Department of Electrical and Computer Engineering, NTUA, Greece
{dimdano,kachris,dsoudris}@microlab.ntua.gr
[2] Democritus University of Thrace, Greece

Abstract. Covid-19 pandemic has devastated social life and damaged the economy of the global population with a constantly increasing number of cases and fatalities each day. A popular and cheap screening method is through chest X-Rays, however it is impossible for every patient with respiratory illness to be tested fast and get quarantined in time. Thus, an automatic approach is needed which is motivated by the efforts of the research community. Specifically, we introduce a Deep Neural Network topology that can classify chest X-Ray images from patients in 3 classes; Covid-19, Viral Pneumonia and Normal. Detecting COVID-19 infections on X-Rays with high accuracy is crucial and can aid doctors in their medical diagnosis. However, there is still enormous data to process which takes up time and computer energy. In this scheme, we take a step further and deploy this Neural Network (NN) on a Xilinx Cloud FPGA platform which as devices are proven to be fast and power efficient. The aim is to have a medical solution on the Cloud for hospitals in order to facilitate the medical diagnosis with accuracy, speed and power efficiency. To the best of our knowledge, this application has not yet been considered for FPGAs while the accuracy and speed achieved surpasses any previous known implementation of NNs for X-Ray Covid detection. Specifically, it can classify X-Ray images at a rate of 3600 FPS with 96.2% accuracy and a speed-up of 3.1× vs GPU, 17.6× vs CPU in performance and 4.6× vs GPU, 13.1× vs CPU in power efficiency.

Keywords: Deep Learning · Neural Networks · Covid-19 · Medical Diagnosis· X-Rays · FPGA · Xilinx

1 Introduction

The sudden spike in the number of patients with COVID-19 , a new respiratory virus, has put unprecedented load over healthcare systems across the world. The COVID-19 pandemic continues to have a devastating effect on the health and economy of the global population.

* Covid4HPC has been open-sourced and is available online in the following Github repository: https://github.com/dimdano/Covid4HPC

A critical step towards the fight against Covid is the effective screening of infected patients so that those infected can receive immediate treatment and get quarantined in time, especially those at high risk. The detection of the disease is commonly performed through chest X-Ray radiograph (CXR) examination by highly-trained specialists. It's a popular and widespread method which usually manifests as an area of increased opacity on CXR. However, it is a very time-consuming, and complicated manual process that can put additional stress on healthcare systems. Also, X-Ray images of pneumonia in the case of Covid-19 are often not very clear and can be misclassified to other diseases or other benign abnormalities which can lead to wrong medication or late quarantine [13, 9].

In this scheme, an automated method is needed to categorize chest X-rays and determine the type of disease. Inspired by the open source efforts of the research community and the urgent need to develop solutions to aid in the fight against the COVID-19 pandemic we will introduce a Deep Learning application for automatic Covid-19 detection through chest X-Rays. The aim of this project is to port our tool in the Cloud so that doctors and clinics from all around the world can have access to a Cloud Medical AI Assistance remotely. Cloud Computing will enable the on-demand availability of computer system resources that will perform Medical Diagnosis using our tool. However, CNN models have high demands in compute power as they are often computationally intensive [7]. Thus, modern companies that operate the data centers will require a demanding workload and computer power due to these algorithms, especially in the pandemic era with terabytes of patient data to be processed everyday.

In this project we take a step further to this modern challenge and provide a solution which, to the best of our knowledge, has not been considered yet. This computational complexity motivated our efforts to enhance these AI task using hardware-specific optimizations by leveraging a hardware architecture known as Field-programmable gate array (FPGA). FPGA-based acceleration has shown great potential [6, 5] as they offload specific tasks from the CPU improving the global performance of the system and reducing its dynamic power consumption. These implementations have seen great advancement on Deep Learning as is it shown that they have been extremely effective on CNN tasks due to their massive parallelism and reconfigurability on the bit level. Thus, by deploying our CNN model to FPGA platforms we accelerated the Image Recognition process with great speed and power efficiency as well, which is essential in datacenter workloads.

In this project, we will describe several highly accurate Deep Learning models using custom and novel Convolutional Neural Network topologies that can detect Covid-19 disease in chest X-Ray images which are also deployed in a Xilinx Cloud FPGA platform. In summary, the main contributions of the paper are as follows:

- We describe three efficient CNN model architectures in terms of memory and size which can classify chest X-Rays in 3 classes (Covid, Viral Pneumonia, Normal). Trained on Tensorflow Deep Learning framework we achieved a maximum of $\sim 97\%$ accuracy which surpasses the performance of previous CNNs for chest X-Ray Covid detection.

- We introduce FPGA-specific optimizations on the model topologies and use 8-bit quantization for the arithmetic precision. We select the most efficient model and accelerate it on a Xilinx Alveo U50 FPGA using an heterogeneous architecture that is also scalable and can be seamlessly ported in datacenters for cloud workoads.
- We ran the FPGA application in a containerized environment and measure the results in terms of accuracy, speed and power efficieny. We outperformed other high performance devices (Xeon CPU, V100 GPU) in performance and performance/watt. Also, we make an evaluation on the model classification ability using heatmaps on top of X-Rays and show other useful classification metrics.

2 Related Work

X-Ray detection algorithms have been explored by many researchers especially in the past, usually for lung diseases such as Viral Pneumonia. In the last year (2020), due to the Covid-19 pandemic there is a continuous growing interest from the research community on developing AI models for the detection of Covid-19 disease. There is an urgent need to aid clinicians in their medical diagnosis using automated tools through Deep Learning. However, as the computational complexity of AI models is large and the laboratory data from new patients grows exponentially a new robust platform is needed to handle all these requests with high speed and efficiency. The following related work involves similar design methods for relatively similar problems or present a related problem that covers our problem domain.

Several studies have investigated the use of Neural Networks towards Covid detetion through chest X-Rays. Mangal et al. [11] presented CovidAID, a deep neural network based model to triage patients for appropriate testing. On the publicly available covid-chestxray dataset their model gave 90.5% accuracy for the COVID-19 infection. On the same scheme, Wang et al. introduced CovidNet [16], a deep convolutional neural network design tailored for the detection of COVID-19 cases from CXR images. They showed a classification report with 93.3% achieved accuracy. Although, these projects were some of the early work towards the fight against Covid and were very useful for the community, they lack the performance of our CNN model which achieves an accuracy of 96.2%.

Also, various deep learning based approaches have been developed to identify Covid-19 [17, 3, 12] but lack the accuracy or precision of our model. Others, such as Jain et al. [10] who presented an Xception model topology for the same problem domain, they achieved slightly higher accuracy as our model (97, 9%) but they do not include a method to accelerate or run more efficiently the inference procedure. Last, there is numerous work that focus on two-class classification between Covid and non-Covid images [14, 8] in contrast with ours which employs a third classification category called Viral Pneumonia. This is essential for the treatment strategy since Viral Pneumonia patients require different treatment plans.

Last, it's worth noticing that to the best of our knowledge there is no previous work on Covid detection through FPGAs. Thus, we will compare some related projects on this problem domain which mainly involves hardware acceleration of CNNs targeting Pneumonia detection which is a fair comparison for our hardware implementation. For example, Chouhan et al. [15] developed a CNN model using the transfer learning approach from models in ImageNet. They reported an average inference computation time of $0.043s$ in an Nvidia GTX 1070 GPU card. Also, Azemin et al. [2] implemented a ResNet-101 CNN model architecture for Covid-19 detection and reported a speed of 453 images/min in CPU.

To conclude, a lot of work has been done using CNNs and Covid-19 detection. Our project presents a novel CNN with higher accuracy than previous known work but also an acceleration method for deployment in cloud FPGAs.

3 Software Implementation

In this study, we constructed several CNN topologies in order to classify the CXR images of the dataset. The best AI model was selected in terms of accuracy and efficiency in order to be deployed in the FPGA (described in next section). In this section, we focus on the analysis of the problem formulation, the dataset used, the training procedure and the hw-oriented optimizations that we did on the models in order to efficiently deploy them on the FPGA device.

3.1 Dataset

We developed the database of Covid-19 X-Ray images from the Italian Society of Medical and Interventional Radiology (SIRM) COVID-19 DATABASE [1, 4]. The number of samples used to train and evaluate the AI models is comprised of a total of 2905 CXR images, split in 219 for Covid, 1345 for Viral Pneumonia and 1341 for Normal class. Although the dataset is relatively small and irregular, we managed to introduce several techniques to overcome this issue as we will describe next.The choice of this dataset was guided by the fact that it is open source and fully accessible to the research community and the general public, and as the datasets grow we will continue to develop and finetune the models accordingly. In the bar chart below, we show the CXR images distribution for each infection type split to train and test sets (test dataset set to 25%).

3.2 Model Topology

We propose three different topologies for this problem in order to have a better evaluation on the dataset and select the most suitable model for acceleration on the FPGA afterwards. We developed separate models that each has different prediction accuracy, architectural complexity (in terms of number of parameters) and computational complexity (in terms of number of MAC operations). A CustomCNN which is a classic convolution neural network, a lightResNet which is a ResNet50 variant and DenseNetX which is based on DenseNet architecture but it also includes the Bottleneck layers and Compression factor.

3.3 Training

Below we will analyze several techniques that we applied on the training procedure. These optimizations mainly had to do with the specific dataset characteristics but also include several hardware-aware optimizations on the model that helped us deploy and accelerate the CNNs in the FPGA more efficiently.

Class weighting Training with a dataset like ours with very few Covid-19 images as opposed to Normal or Viral Pneumonia images constitutes a class-imbalanced problem. This is a complexity that poses significant challenge to the converging of our models as CNNs are normally assume to be trained on identical distribution datasets. To overcome the class variance we imposed specific class weights (i.e. 6× on Covid class) which applied to the model's loss for each sample and eventually helped the model learn from the imbalanced data.

HW-aware optimizations The CNNs' topology needed some minor modifications in order to be compatible and efficient with Vitis AI quantizer and compiler. In particular, the order of the Batch Normalization (BN), Rectified Linear Unit (ReLU) activation and Convolution layers has been altered from BN⇒ReLU⇒Conv to Conv⇒BN⇒ReLU. Also, another optimization that we did is in the case of GlobalAveragePooling2D, which we needed for example in DenseNetX and we replaced it with AveragePooling2D plus a Flatten layer. Last, softmax was implemented in the DPU and not in SW (proved to be more than 100 times faster) using an AXI master interface named `SFM_M_AXI` and an interrupt port named `sfm_interrupt`. The softmax module used `m_axi_dpu_aclk` as the AXI clock for `SFM_M_AXI` as well as for computation.

4 Hardware and System Design

In this section we will describe the process of quantizing, evaluating, compiling and at last running the AI models on the Alveo FPGA platform. Also, we analyze the full architecture of the FPGA design working in an heterogeneous system that allows efficient communication with the host processor. Last, we create a full end-to-end environment that anyone can use and test our project seamlessly through an FPGA-containerized application.

4.1 Acceleration Method

As our application can be used from millions of users across the globe, an efficient and fast solution is needed. We chose to use Vitis AI environment in order to deploy our CNN models in an Xilinx Alveo U50 FPGA and eventually create an application with high inference throughput and small memory footprint as well, an essential factor for cloud workloads.
Below, we analyze further the steps required for model quantization and evaluation and last the compilation of the model which creates the DPU instructions for utilizing the compute units (CUs) of the FPGA.

1. *Quantization*: First, before proceeding with the quantization process we converted our models to a Tensorflow compatible floating-point frozen graph. Next, we chose to quantize the trained weights of our CNNs with 8-bit precision as this has been proven to keep an acceptable accuracy at similar CNN applications. Last, we provided a sample set of the training data to calibrate the quantization process. The data performed a full forward pass through the model and the weights were calibrated according to the data range the application needs for inference.
2. *Evaluation of quantized model*: The conversion from a floating-point model where the values can have a very wide dynamic range to an 8-bit model where values can only have one of 256 values almost inevitably leads to a small loss of accuracy. Thus, the prior evaluation of the quantized graph on Tensorflow was essential before proceeding with the compilation of the model. Nevertheless, this technique most of the times gave almost identical accuracy results compared with the actual application tested on the board. Also, the quantized graph on the FPGA compared with the floating point graph on CPU had a small impact on the final accuracy ($< 0.5\%$).
3. *Model compilation*: In the final stage, we compiled the graph into a set of micro-instructions that were passed to the DPU in `.xmodel` file format. The Vitis AI compiler converted and optimized where possible the quantized deployment model and gave as output the final "executable" for CNN inference. The generated instructions were specific to the particular configuration of our DPU which in our case the `DPUCAHX8H` DPU IP was selected. We passed the DPU's parameters in a `.dcf` file for the target Alveo U50 board.

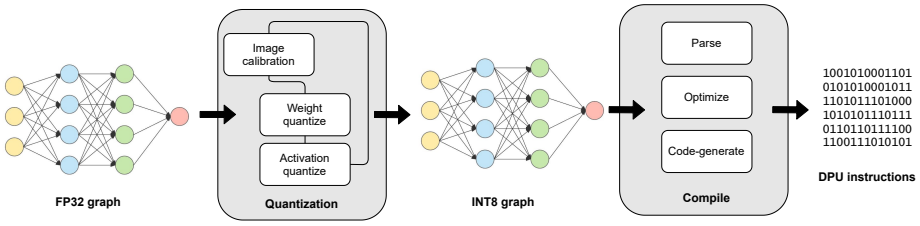

Fig. 1. CNN graph quantization and compilation for the FPGA DPU

5 Evaluation and Results

In this section we will evaluate and profile our application. We will start from the model performance of the neural networks in terms of validation loss, accuracy and several other classification metrics. Next, we will proceed with the performance evaluation of the hardware accelerator in terms of resource utilization, acceleration and power efficiency achieved over CPU and GPU. Last, we will present some qualitative results that help localizing areas in the X-Ray image most indicative of Covid or Viral Pneumonia.

5.1 Model evaluation

In this study, experiments were ran with Tensorflow and Keras using the typical 224×224 image dimensions found in most CNNs. All the models have been trained with Adam optimizer along with EarlyStopping and best model callbacks. The classification models were optimized by minimizing the cross-entropy loss function. Also, several parameters and hyperparameters of each model were tuned during the training such as Learning Rate (LR) and epochs. Below, in Table 1 you can observe the major characteristics of each model in terms of training hyperparameters, model specifications and model evaluation.

Table 1. CNN model characteristics and performance

	Hyperparameters		**Model specs**		**Evaluation**	
Model	***LR***	***Epochs***	***Params***	***FLOPs***	***Accuracy***	***Loss***
CustomCNN	0.0001	70	2.033G	1.025G	96.2%	0.16
lightResNet	0.001	60	2.697G	2.814G	96.5%	0.408
DenseNetX	0.005	80	0.758G	1.722G	94.9%	0.264

5.2 Qualitative analysis

In the previous tables we showed a variety of performance criteria that can be used to evaluate the performance of our classification models. Under certain circumstances we can select a different model with certain characteristics that satisfies our needs depending on the performance efficiency (FLOPs) or accuracy. For demonstration purposes and the FPGA implementation we selected the CustomCNN model which has the most efficient performance as it achieves high accuracy (almost same as lightResNet) while having minimal compute requirements which is essential for compute workloads. Additionally, below we present several useful activation maps that were obtained for the last convolutional layer of CustomCNN network. These are very important because they give us an insight on the model's classifier capability as well as validate the regions of attention of the disease.

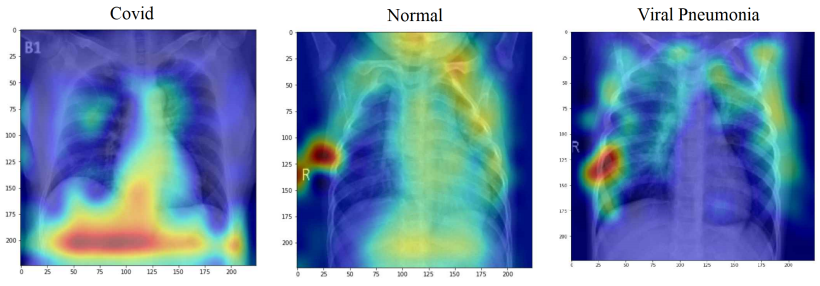

Fig. 2. X-Ray visualizations using attention heatmaps

5.3 System performance

For the evaluation of the system design we first confirmed the resource utilization of the FPGA DPU. The hardware setup for the deployment was a Xilinx Alveo U50 Cloud FPGA with 8GB HBM Memory Capacity and 316 GB/s total bandwidth. The device was installed on Gen4x8 PCI express running at a kernel clock of 300MHz. Table 2 shows the resource utilization of a DPUv3E kernel with five batch engines (our design used two kernels).

Table 2. Resource utilization of a single DPU kernel

	Utilization summary				
Name	***BRAM***	***URAM***	***DSP***	***FF***	***LUT***
Used	628	320	2600	310752	250290
Percentage	46.7%	50%	43.6%	21.2%	28.7%

Next, we evaluated inference with the CustomCNN model on other high performance systems, specifically an Nvidia V100 GPU and a 10-core Intel Xeon Silver 4210. The inference on the other devices was tested on Tensorflow with the default settings and choosing appropriate batch sizes. On the left side of Figure 3 we can observe the maximum throughput each device achieves measured with X-Rays per second (FPGA: 3600, GPU: 1157, CPU: 204). Also, we annotated the latency (in ms) for single X-Ray image inference in each device. Additionally, on the right side of Figure 3 we show the measurement for the power efficiency of each device in X-Rays/Sec/Watt (FPGA: 51.3, GPU: 11.1, CPU: 3.9).

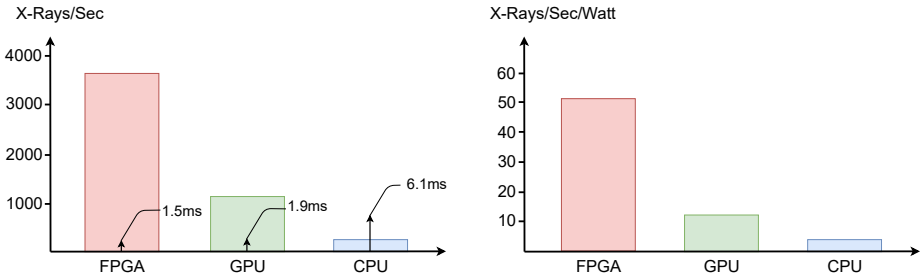

Fig. 3. Performance and Performance/Watt metrics across different architectures

The throughput metric is essential for cloud workloads that operate with large size of patient data while the latency metric is important for edge scenarios (i.e. mobile phones) which are time critical and an immediate response is needed.

It is evident from the quantitative comparison that the FPGA exhibits the highest inference speed for large batch size scenarios in the cloud achieving 3.1× speed-up from GPU and 17.6× speed-up from CPU in throughput. It's worth mentioning that the 8-bit quantized CNN used on FPGA had less than 0.5% accuracy loss which was desired for the performance and power efficiency gains.

Last, the FPGA outperformed the other two devices in the power efficiency metric measured in X-Rays/Sec/Watt. Specifically, it achieved 4.6× speed-up from GPU and 13.1× speed-up from CPU. The 51.3 X-Ray/Sec/Watt metric of the FPGA means that in order to specify a disease in a single chest X-Ray image we would only need $0.019s$ and 1 Watt of compute power. This is crucial for cloud providers which aim to reduce energy consumption in datacenters while maintaining the required performance for the needs of each application.

6 Conclusion

In this study, we introduced several AI models each with its own characteristics for the detection of COVID-19 cases from CXR images that are open source and available to the general public. We demonstrated significant improvement in accuracy and performance compared with the aforementioned related work. Moreover, we investigated how our model makes predictions using an attention heatmap method in an attempt to gain deeper insights into critical factors associated with COVID cases, which can aid clinicians in improved screening as well as improve trust and transparency when leveraging our CNN.

Moreover, we quantized, compiled and accelerated the AI model for deployment on an Alveo U50 FPGA with the aim to accelerate the computer-aided screening. The application was containerized and can be seamlessly ported in a cluster of FPGAs operating in the cloud with high performance and energy efficiency compared with other architectures.

By no means this is a production-ready solution intended for self-diagnosis. From a research point of view, we are focusing on boosting more the performance and adding further features in our AI Health framework as new data is collected such as risk stratification for survival analysis or predicting hospitalization durations. The spectrum of possible AI automated systems is vast but this work shed some light to the area with successful results aiming to make FPGAs contribute fundamentally into the computer-aided Medical Diagnosis.

Acknowledgements This project was funded from the Xilinx University program and the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 2212- Hardware Acceleration of Machine Learning Applications in the Cloud.

References

1. COVID-19 DATABASE. https://www.sirm.org/category/senza-categoria/covid-19/, accessed: 2020-10-30
2. Azemin, M., Hassan, R., Mohd Tamrin, M.I., Ali, M.: Covid-19 deep learning prediction model using publicly available radiologist-adjudicated chest x-ray images as training data: Preliminary findings. International Journal of Biomedical Imaging **2020**, 1–7 (08 2020). https://doi.org/10.1155/2020/8828855

3. Channa, A., Popescu, N., Malik, N.U.R.: Robust technique to detect covid-19 using chest x-ray images. pp. 1–6 (10 2020). https://doi.org/10.1109/EHB50910.2020.9280216
4. Chowdhury, M., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M., Mahbub, Z., Islam, K., Khan, M.S., Iqbal, A., Al-Emadi, N., Reaz, M.B.I., Islam, M.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665 –132676 (07 2020). https://doi.org/10.1109/ACCESS.2020.3010287
5. Danopoulos, D., Kachris, C., Soudris, D.: Acceleration of image classification with caffe framework using fpga. pp. 1–4 (05 2018). https://doi.org/10.1109/MOCAST.2018.8376580
6. Danopoulos, D., Kachris, C., Soudris, D.: Automatic generation of fpga kernels from open format cnn models. pp. 237–237 (05 2020). https://doi.org/10.1109/FCCM48280.2020.00070
7. Danopoulos, D., Kachris, C., Soudris, D.: Utilizing cloud fpgas towards the open neural network standard. Sustainable Computing: Informatics and Systems **30**, 100520 (2021). https://doi.org/https://doi.org/10.1016/j.suscom.2021.100520, https://www.sciencedirect.com/science/article/pii/S2210537921000135
8. Dansana, D., Kumar, R., Bhattacharjee, A., D, J., Gupta, D., Khanna, A., Castillo, O.: Early diagnosis of covid-19-affected patients based on x-ray and computed tomography images using deep learning algorithm. Soft Computing (08 2020). https://doi.org/10.1007/s00500-020-05275-y
9. DAVIES, H., DELE MD, M., E.-L., E.: Reliability of the chest radiograph in the diagnosis of lower respiratory infections in young children, the pediatric infectious disease journal. The Pediatric Infectious Disease Journal **15**, 600–604 (1996)
10. Jain, R., Gupta, M., Taneja, S., D, J.: Deep learning based detection and analysis of covid-19 on chest x-ray images. Applied Intelligence pp. 1–11 (10 2020). https://doi.org/10.1007/s10489-020-01902-1
11. Mangal, A., Kalia, S., Rajgopal, H., Rangarajan, K., Namboodiri, V., Banerjee, S., Arora, C.: Covidaid: Covid-19 detection using chest x-ray (04 2020)
12. Miranda Pereira, R., Bertolini, D., Teixeira, L., Silla, C., Costa, Y.: Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. Computer Methods and Programs in Biomedicine **194**, 105532 (05 2020). https://doi.org/10.1016/j.cmpb.2020.105532
13. Neuman, M.I., Lee, E.Y., Bixby, S., Diperna, S., Hellinger, J., Markowitz, R., Servaes, S., Monuteaux, M.C., Shah, S.S.: Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. Journal of Hospital Medicine **7**(4), 294–298 (2012)
14. Pham, T.: A comprehensive study on classification of covid-19 on computed tomography with pretrained convolutional neural networks. Scientific Reports **10** (10 2020). https://doi.org/10.1038/s41598-020-74164-z
15. Singh, S., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damasevicius, R., Albuquerque, V.: A novel transfer learning based approach for pneumonia detection in chest x-ray images. Applied Sciences **10**, 559 (01 2020). https://doi.org/10.3390/app10020559
16. Wang, L., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images (03 2020)
17. Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., Xia, Y.: Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. IEEE Transactions on Medical Imaging **PP**, 1–1 (11 2020). https://doi.org/10.1109/TMI.2020.3040950