Covid-19 Lung Disease Detection from X-Ray Images

Elizaveta German
elizaveta.german@studenti.unipd.it

Andrii Kliachkin
andrii.kliachkin@studenti.unipd.it

***Abstract*—This paper aims to introduce a machine learning technique based on convolutional neural networks and the CNN attention mechanism to automatically detect Covid-19 cases from chest X-ray (CXR) images. Three different experiments were conducted: testing different neural network architectures, trying various image contrasting techniques, and adding an attention mechanism to the most performative model. The best accuracy was achieved by AlexNet model with 93,4% on test set. Image preprocessing with histogram equalization and adaptive equalization has also increased the performance of the model. The attention mechanism based on AlexNet outperformed the previous model and allowed to enhance True Positive Rate from 94% up to 98,6% thereby minimizing the percentage of missed cases, although increasing False Positive Rate. Another contribution is further research into the effects of the CNN attention mechanism on model performance, which is a relatively overlooked field.**

I. Introduction

Since the onset of the COVID-19 pandemic, there have been numerous efforts to automate the process of detection of infected patients using deep learning techniques. These started with various combinations of traditional image classification models, mainly CNNs [1]. Later, unorthodox and entirely new neural network architectures were tried, such as COVID-Net [2]. Most of these new models, however, are extremely complicated and computationally taxing. This paper aims to achieve similar performance with limited resources by combining reasonably "lightweight" models with preprocessing techniques and architectural designs, which, to the authors' best knowledge, were not yet applied to the task of COVID-19 diagnostics.

The rest of this paper is organized as follows. In Section III we present an overview of the data and the technologies that will be used. In Section IV the input data and preprocessing performed to it is analysed more thoroughly. In Section V, we describe in greater detail the models used in the experiments and their initialization, as well as metrics used to evaluate their performance. The attention mechanism is also described in this section. Outcomes of experiments are presented in Section VI, where we find that two best-performing architectures are AlexNet with no attention module and AlexNet with attention added after the third convolution layer. Authors' conclusions are presented in Section VII.

II. Related Work

As stated earlier, the hunt for better performance in deep learning-assisted COVID diagnostics started with unmodified well-established CNN architectures [1], [3]. While these and other early studies achieved impressive results, they were typically conducted on relatively small or private datasets, harming both performance and reproducibility. Furthermore, they focused primarily on Computed Tomography (CT) imaging, which was eventually largely phased out in favour of Chest X-Ray (CXR) imaging for practical concerns (such as availability of medical imaging equipment [4]), especially after large public CXR datasets were constructed, i.e., [2], [5].

COVID-Net [2], a neural network specifically tailored for COVID detection from CXR images through a human-machine collaborative design strategy, signified a large advance in the field with accuracy of 93.3% and a sensitivity of 91%. In this article, an attempt will be made to surpass this value without utilizing pre-training.

Cohen et al. [6] were the first to utilize a novel DenseNet model [7] to fight the COVID pandemic. This architecture, which utilizes a set of residual connections between each convolutional block, has achieved high accuracy in predicting coronavirus pneumonia severity; however, it was not tasked with detecting the presence of a COVID infection.

III. Processing Pipeline

Dataset. The dataset contains 4575 chest X-ray posteroanterior (taken from the patient's back) images divided into three groups by the patient's condition - normal, COVID-19 and non-COVID pneumonia, with 1525 images for each case. The dataset was downloaded from Mendeley Data [5].

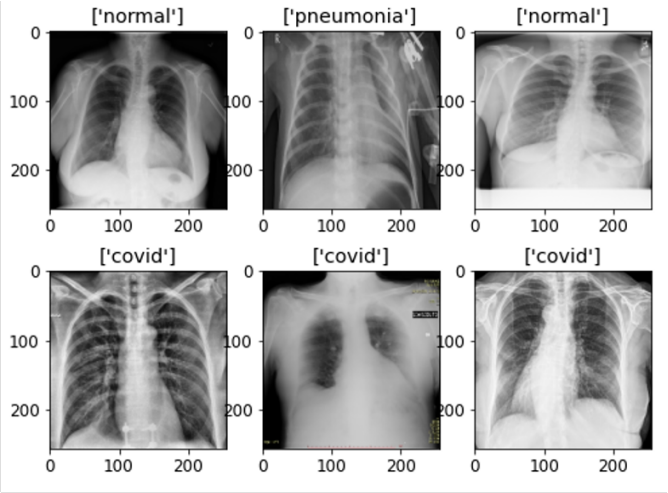

Fig. 1: Example of X-Ray images

Objective. The aim of the current research is to detect COVID-19 disease from X-ray images and to minimize the percentage of false negative predictions. For the classification of lung conditions, the three main experiments are conducted in order to find the best performative model. Firstly, different architectures of convolutional neural networks for image classification are tested. After that, the most performative architecture is used with three various types of image contrasting. Finally, the attention mechanism is applied, aiming to improve model performance.

Convolutional Neural Networks. CNN is a multilayer neural network where neurons have local receptive fields allowing to extract features at each layer of the network such as horizontal edges. The deeper is the CNN, the more complex features it can identify - for example, faces or objects. The basic CNN architecture consists of a set of convolutional, pooling and fully-connected layers.

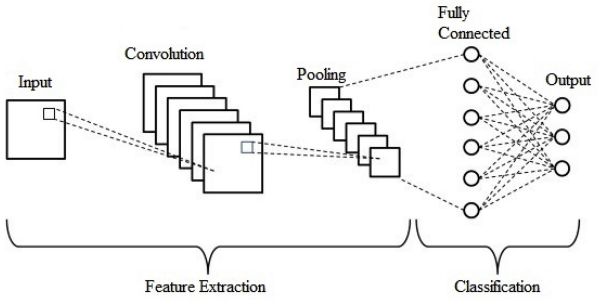

Fig. 2: A typical CNN architecture [8]

Image Contrasting Techniques. Contrasting is a preprocessing technique that takes the distribution of pixel densities and stretch it to a wider range of values, increasing the level of contrast between the lightest and darkest parts of the image. Three different approaches are tested - contrast stretching, histogram equalization and adaptive histogram equalization. Examples are depicted in Fig.3, Fig.4 and Fig.5, respectively.

The contrast stretching attempts to improve the contrast by rescaling the intensity values to the full range of pixel values possible.

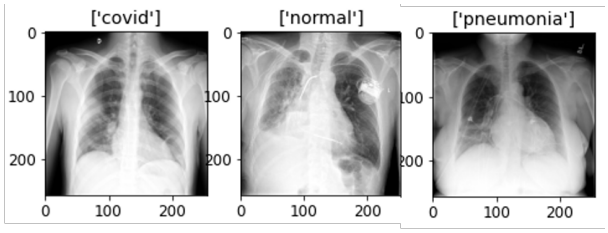

Fig. 3: Contrast Stretching

The histogram equalization enhances the image contrast by plotting pixel intensities on a histogram and spreading out the most frequent intensity values to cover the areas with lower contrast, so that they are gaining higher contrast.

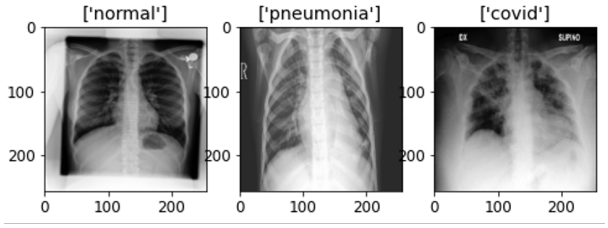

Fig. 4: Histogram Equalization

In the adaptive histogram equalization, several histograms are computed instead of one, where each corresponds to a different section of the image. It is an effective method of contrasting, but is more prone to overamplifying noise in some regions of the image. A version of adaptive equalization called Contrast Limited Adaptive Histogram Equalization (CLAHE) aims to prevent the overamplification.

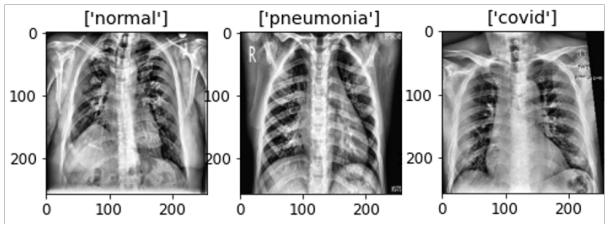

Fig. 5: Adaptive Histogram Equalization

Attention Mechanism. The attention mechanism, as described in [9], aims to increase the performance of a CNN by extracting feature maps, each of which forms an intermediate representation of the input image, from different stages in the network pipeline and forcing the model to make a classification decision based solely on a weighed combination of these maps. This amplifies the influence of important regions of the image, while suppressing irrelevant ones.

Experimental Setup. The overall system for detection of COVID-19 disease consists of several parts. First, raw images of different resolution were resized into a uniform shape 256 x 256 and then normalized. The dataset was split into training, validation and test subsets, and then several models with different architectures were trained. The models were evaluated by accuracy, recall/precision matrices and F1-score. The effect of data augmentation was also tested. The model complexity was assessed by number of parameters, memory usage and time required for the training. After that, the most performative model was chosen in order to experiment with different contrasting techniques and attention mechanism and to evaluate their effect on the performance of the model.

IV. Signals and Features

The input images are X-ray scans of different image resolution. The first step of processing data is resizing to the uniform shape 256x256, that is a suitable size from computational point of view. The lower resolution will result in accuracy decay. After resizing the images were transformed into arrays where each element represents the pixel intensity of the image.

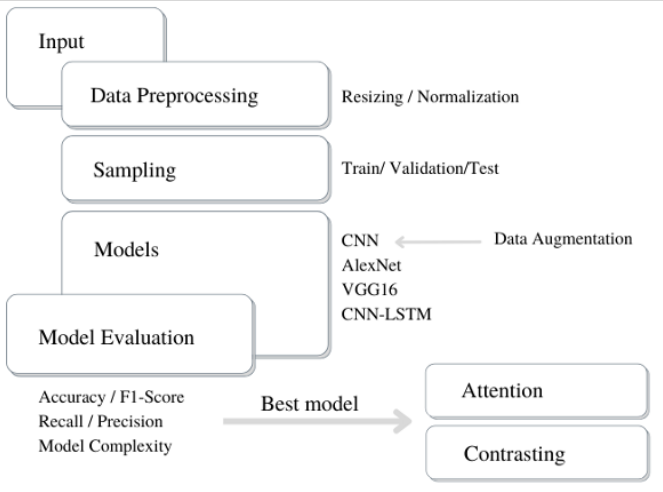

Fig. 6: The overall system of COVID-19 Disease Detection

Then image pixels were normalized so that the range is from 0 to 256. The input images were RGB.

The classes of the images were assigned according to the categorization: 0 - for normal condition, 1 - for COVID disease and 2 - for pneumonia. The balanced data (1525 for each class) of 4575 samples were split into training, validation, and test sets with proportion of 70, 15, 15, respectively, with 3203 samples for training set, 686 for the validation and 686 for the test set.

Data Augmentation is a common way to generate more data for training what can be useful, when a dataset is small. Some of the data used in this research were already augmented, in particularly, the COVID-19 X-ray images [5]. Adding more augmented data was a part of experiments in order to see its regularization effect and how it may or may not improve the model performance. The augmentation includes horizontal flip, shifting width and height to a factor of up to 0.1, a zoom range of up to 0.1, and a rotation range of 10.

While in some other works feature extraction is performed with a pre-trained CNN, in this paper no pre-training is being done. In all experiments in this paper, both feature extraction and classification are performed by different layers of a self-contained network.

V. Learning Framework

Models. The architectures chosen for the task were VGG16 [10], AlexNet [11] and a novel CNN-LSTM [8]. In the case of VGG16 and CNN-LSTM networks, the number of neurons in the final fully connected layers was reduced to accommodate for limited memory and processing capacity. The models were compared with each other and with a baseline shallow CNN using metrics discussed at the end of this section.

The baseline model is a shallow network with only one convolutional and one pooling layer. This model was built as a starting point of the experiments in order to see which performance results may be achieved with the most simple model. The efficiency of data augmentation was tested on this model.

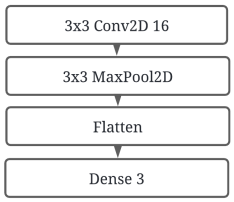

Fig. 7: Baseline model

The next tested architecture is AlexNet with 5 convolutional layers with combination of max pooling layers and three fully connected layers followed by Dropout with 0.5 rate for regularization. The AlexNet was originally used on large-scale Imagenet for prediction of 1000 classes. The last layer was modified in accordance to number of classes in this research.

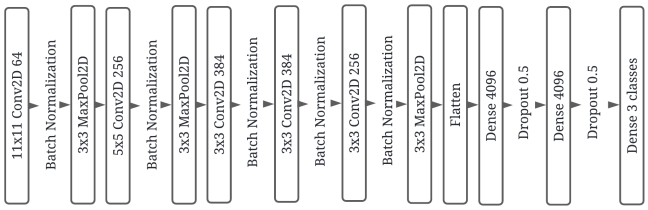

Fig. 8: AlexNet architecture

VGG16 is an architecture with 16 layers - 13 convolutional and 3 fully connected. This is the most used architecture as a pre-trained model in transfer learning. Nevertheless, transfer learning is beyond the scope of current research, and VGG16 was trained manually. In order to fit computational limitation the number of neurons in the last dense layers was reduced to 64.

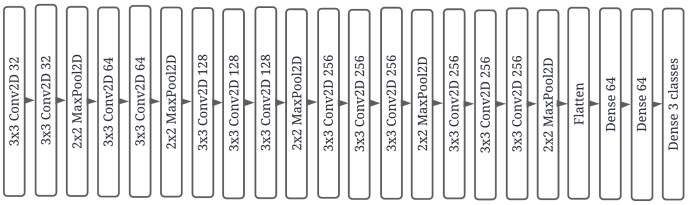

Fig. 9: VGG16 architecture

In the CNN-LSTM architecture, a CNN with 5 convolutional layers is used for feature extraction, and an LSTM layer, followed by two dense layers, are tasked with classification. This model was significantly simplified in comparison to the original one because of limited computational capacity. The original model has been claimed to achieve an accuracy of 97% and a specificity of 100% on a COVID-19 detection task [8].

After choosing the most performative model out of the ones tested, the attention mechanism was added with the aim to enhance accuracy.

Fig. 10: CNN-LSTM architecture

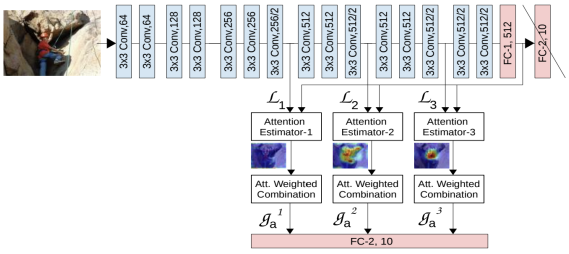

Fig. 11: The attention mechanism applied to an example convolutional neural network [9].

This mechanism is illustrated in Fig. 11. First, sets of feature vectors $L^s = \{l_1^s, l_2^s, \ldots, l_n^s\}$ are extracted at a given convolutional layer $s \in \{1, 2, \ldots, S\}$, where each l_i^s is a vector of activations of the spatial location i of this layer. Then, compatibility scores $\{c_1^s, c_2^s, \ldots, c_n^s\}$ are calculated using a compatibility function $C(\hat{L}^s, g)$ (either weighed compatibility, where $c_i^s = \langle u_i, l_i^s + g \rangle$, and u_i are weights to be learned during model training, or non-parametric dot-product compatibility, where $c_i^s = \langle g, l_i^s \rangle$; the latter one is used in this work). Here g is the global feature vector (the final image representation obtained by the CNN), and $\hat{L}^s$ is the image of L^s under a linear mapping of l_i^s to the dimension of g.

The scores are passed through a softmax function, producing normalized compatibility scores $\{a_1^s, a_2^s, \ldots, a_n^s\}$, which are then used to compute a vector $g_a^s = \sum_{i=1}^{n} a_i^s \cdot l_i^s$ for each layer s. The new attention-incorporating global feature vector is defined as $g_a = [g_a^1, g_a^2, \ldots, g_a^S]$. This vector now replaces g as the global feature vector and is exclusively used for classification.

This approach has been shown to significantly increase the performance of CNNs in generic image classification tasks, but has not yet been applied to the field of COVID-19 detection.

Model Initialization. For all models described above, the initialization parameters were the same, except for varied optimizer and learning rate. Four types of callbacks were used:

1) Time callback in order to track the training time;
2) Early stopping on validation loss with patience 3 and minimum delta 0.0005 for regularization;
3) Reducing learning rate on Plateau on validation accuracy with patience 2 and factor 0.2;
4) Model checkpoint on validation accuracy in order to save weights reached the best performance so far;

For all models, the sparse categorical cross entropy loss function was used, which is a suitable loss function for multilabel classification, which in this case were 3 categories from 0 to 2.

Adam and Stochastic Gradient Descent optimizers were tested, and for each model the best performative one on the test accuracy was chosen. The same was done with learning rate, 0.001 and 0.0001.

Performance metrics. For measuring performance of the models four evaluation metrics were used:

1) Accuracy - the total number of correct predictions;
2) Recall - true positive rate is used to measure the percentage of actual positives which are correctly identified;
3) Precision - positive predictive value is used to measure how precise is a model in predicting positives;
4) F1-Score - the harmonic mean between precision and recall;

The confusion matrix and built on it the recall and precision matrices were also used in order to assess the quality of prediction for each class. Precision is a good measure when the higher cost is associated with false positive predictions. Recall is the most appropriate when higher risk is related to false negative results that is the case of COVID-19 detection (predicted as normal or pneumonia, when it is actually COVID). Since the objective is minimizing the amount of false negative predictions, the recall score should be the main adjudicative metric.

VI. Results

As a first experiment, the comparison of the described above architectures was conducted, and the best performative model was chosen for the further processing. Then three contrasting techniques were tested on that model without modifying the architecture and initialization parameters. And finally, the attention mechanism was applied.

Architecture comparison. Five models were trained - the baseline, the baseline with augmentation, AlexNet, VGG16 and CNN-LSTM. The training and validation accuracy curves of the last three models are presented in Fig.12. As can be seen, overfitting was not an issue in this case, since regularization was already included in the architectures, and it was significantly enough for models to generalize.

The performance results of this experiment can be seen in Fig.13, it shows the metrics used for the evaluation of all five models.

As expected, the more complex architectures AlexNet and VGG16 outperformed the baseline model, however the latter showed the high level of overall accuracy and recall - over 90%. Data augmentation made the results of the base model worse, and the following training was conducted without augmented data. CNN-LSTM showed less accuracy, so the adjudication was between AlexNet and VGG16 models.

The recall and precision matrices of these models are presented in Fig.14 and Fig.15. AlexNet is performing a bit

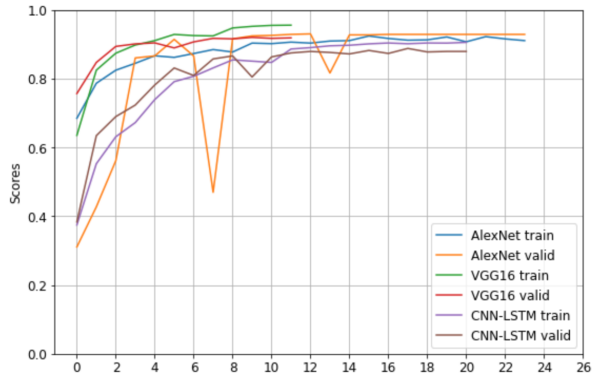

Fig. 12: Training and validation accuracy curves

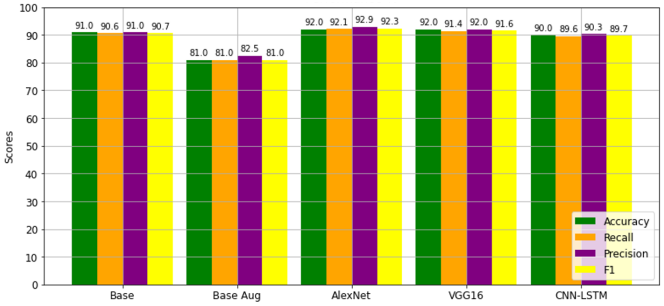

Fig. 13: Performance metrics

more accurate than VGG16, and in particular, the recall. 95% of COVID cases are predicted correctly, and false negative results constitute only 5% of COVID samples - 4% mistakenly predicted as normal, and only 1% predicted as pneumonia.

The worst predicted class in both models is pneumonia with only 86% of correctly identified cases. 11% are predicted as normal while being pneumonia, that can be possibly explained by the initial phase of the disease so that it is too difficult for the models to notice the changes in the X-ray images of lungs. Thus, the models can be less efficient for the cases where pneumonia is in the first stage of developing. Nevertheless, the focus of this research is on COVID-19 detection, and the discussed models provide a sufficient level of accuracy in this scope.

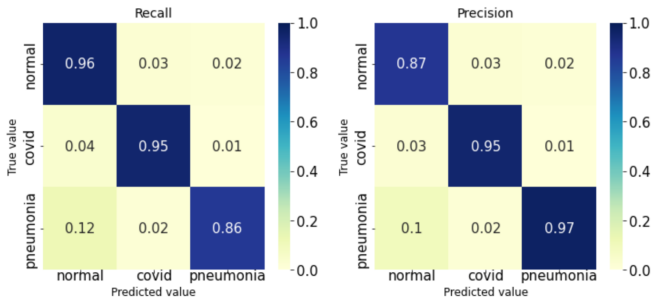

Fig. 14: Recall and Precision matrices of AlexNet

Model complexity. In addition to the performance assessment, the temporal and memory complexity of all models

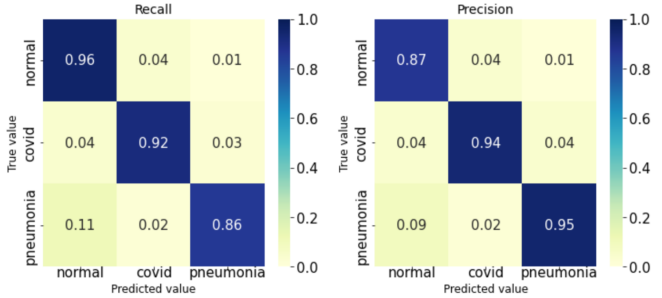

Fig. 15: Recall and Precision matrices of VGG16

was analysed. The running time of the models required for the training can be seen in Fig.16. The training of VGG16 lasted the longest, even if the number of parameters in fully connected layers was considerably reduced. The second longest is AlexNet, the best performative model in terms of accuracy. Time values for the baseline model with and without data augmentation differ due to additional images generated by the process that increased the running time.

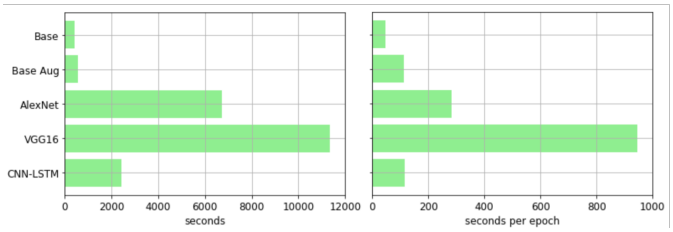

Fig. 16: Total running time and running time per epoch

Memory complexity displays how many MB the model and its weights take up when saved to memory. The comparison can be seen in Fig. 17. AlexNet is the most costly model out of all in terms of memory and with the largest number of parameters. However, it is trained faster, than VGG16, and results in better performance in comparison to other models. Thus, next experiments are based on AlexNet.

Model	Memory, MB	# of parameters
CNN-LSTM	2.1	131 555
Base/Augmented	3.2	786 883
VGG16	19.1	4 733 155
AlexNet	233.3	58 299 139

Fig. 17: Memory complexity

Image Contrasting comparison. AlexNet was trained three times on the same data, but preprocessed with different types of contrasting. The performance results are shown in Fig. 18.

The contrast stretching gave lower accuracy than the model without preprocessing. Histogram and adaptive histogram equalization led to higher performance, increasing accuracy up to 93%. The higher recall score is given by histogram

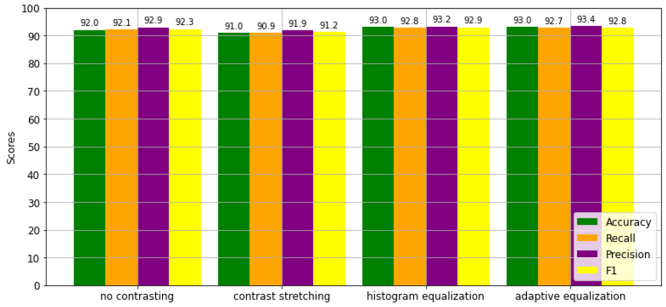

Fig. 18: Contrasting comparison on AlexNet

equalization. By analysing recall matrices it can be seen that both preprocessing techniques give almost the same result in terms of false negative predictions of COVID-19 disease with 2% of COVID cases predicted as normal for histogram equalization and 3% for adaptive equalization.

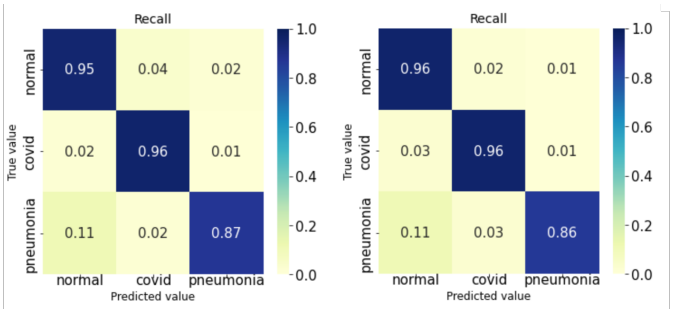

Fig. 19: Recall matrices with histogram equalization (left) and adaptive equalization (right)

The results of the experiment with different contrasting techniques show that it can enhance the performance of the models.

Adding Attention.

As evident from the nature of the attention module, it can be introduced into a CNN in different ways by changing the set of convolutional layers S at which the intermediate features are extracted. In this paper, the following configurations are tried: $S_1 = \{2, 4, 5\}; S_2 = \{2, 4\}; S_3 = \{3\}$. The motivation

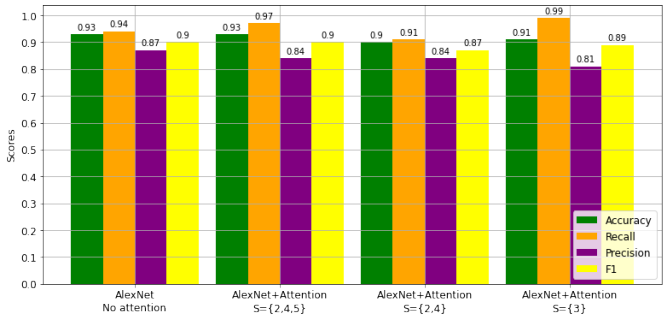

Fig. 20: Performance comparison between AlexNet (FC-256 instead of FC-4096-1; FC-4096-2 removed),without the attention mechanism and the same network with different attention configurations

for these choices stem from the conclusions of the original paper - namely, that the features extracted should be relatively "mature" - and from AlexNet architecture, which is relatively shallow at just 5 convolutional layers. One of two final FC layers was eliminated to accommodate for limited processing power, and the number of neurons in the second one was reduced from 4096 to 256, which didn't result in any significant decrease in performance. Experiment results are illustrated in Fig. 20.

While the accuracy remained mostly unchanged with the introduction of attention, and even declined by 3% for S_2 and 2% for S_3, recall score increased from 0.93 to 0.986 for S_3 - but at the cost of precision.

It should be noted that although the number of parameters and memory complexity remains relatively unchanged with difference of at most 5% between attention-incorporating models, it is significantly reduced in comparison to the original AlexNet model due to the elimination of one FC layer.

Fig. 21 shows a Grad-CAM [12] visualization of the regions of an input image treated by the neural network as important for the final classification decision.

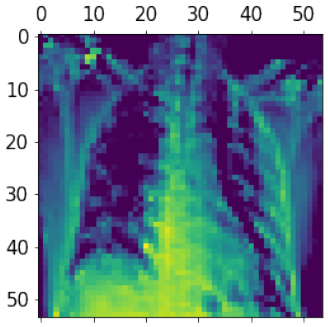

Fig. 21: A Grad-CAM visualization of the loss function gradient on an X-ray of a COVID-positive patient

VII. Concluding Remarks

The performance of four different CNN models in COVID detection on X-ray images has been evaluated. The best-performing model was picked and outfitted with an attention mechanism with the aim of further performance increase. Adding data augmentation resulted in decrease of accuracy and longer training time, and the further modeling was done without augmented data.

The best accuracy on test set (93.4%) was achieved by unaugmented AlexNet, while the highest recall - arguably, the most important metric in the given task - was achieved by AlexNet with attention introduced after the third convolution layer. This is, however, coincided with a fall in precision - which suggests that the mechanism has just shifted the model bias towards the positive (COVID) class. Although high True Positive Rate is an extremely important quality in disease detection systems, this effect may be negated by an increase in False Positive Rate, which can lead to unneeded

hospitalization of COVID-negative patients and increase the already high strain on healthcare systems.

The most important limiting factor during the course of this study has been computing power. One of the possible directions of future works is conducting a similar experiment on more powerful equipment - this would facilitate a more balanced evaluation of network performance without simplifying the architecture and allow for a full-scale testing of the attention mechanism instead of a proof-of-concept presented in this paper; however, it is likely that introducing the mechanism into a more sophisticated network would bring about the same results as presented here.

REFERENCES

[1] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su, G. Lang, Y. Li, H. Zhao, J. Liu, K. Xu, L. Ruan, J. Sheng, Y. Qiu, W. Wu, T. Liang, and L. Li, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020.

[2] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Scientific Reports*, vol. 10, Nov. 2020.

[3] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid ai development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection patient monitoring using deep learning CT image analysis," 2020.

[4] E. J. van Beek, S. Mirsadraee, and J. T. Murchison, "Lung cancer screening: Computed tomography or chest radiographs?," *World Journal of Radiology*, vol. 7, pp. 189–193, Aug 2015.

[5] A. Asraf and Z. Islam, "COVID19, pneumonia and normal chest X-ray PA dataset," 2021.

[6] J. P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A. F. Abbasi, B. Shen, H. K. Mahsa, M. Ghassemi, H. Li, and T. Q. Duong, "Predicting COVID-19 Pneumonia Severity on Chest X-ray With Deep Learning," *Cureus*, vol. 12, p. e9448, Jul 2020.

[7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[8] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in medicine unlocked*, vol. 20, p. 100412, 2020.

[9] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, p. 336â€“359, Oct 2019.